Exploring methods for health-related quality-of-life instrument translation and validation:

A first look at the Norwegian EORTC-QLQ-LMC21

Bethany Kirsten Danielsen

Thesis submitted as a part of the Master of Philosophy Degree
in Health Economics, Policy and Management

Department of Health Management and Health Economics
The Faculty of Medicine

UNIVERSITY OF OSLO

May 2015

2015

Exploring methods for health-related quality-of-life instrument translation and validation - A first look at the Norwegian EORTC-QLQ-LMC21

http://www.duo.uio.no/

Print: Reprosentralen, University of Oslo

Abstract

BACKGROUND: Colorectal cancer is the second most common cancer in Norway, with the incidence rate increasing for both men and women. Patients with colorectal cancer commonly experience metastases in the liver. In Norway, about 30% of patients will have liver metastases at the time of diagnosis, and another 20% will develop metastases during the course of their treatment. Because interventions aimed at curing or managing this disease may have a negative impact on patient health, measuring patient outcomes in the form of health-related quality-of-life is important to assess the relative benefits of these interventions to patients. Disease specific health-related quality-of-life measures have been found to be more sensitive to the health states of patients, however the disease specific measure used for assessing health-related quality-of-life in colorectal cancer patients with liver metastases has previously not been available in the Norwegian language.

OBJECTIVE: This cross-sectional study seeks to explore the methods used in the translation and psychometric assessment of health-related quality-of-life instruments and preliminarily assess the quality of the Norwegian EORTC-QLQ-LMC21 in terms of validity, reliability, responsiveness and equivalence with the English version.

METHOD: This study is divided into two parts: (1) the qualitative translation process; and (2) the qualitative psychometric assessment of the resulting translated instrument. The EORTC-QLQ-LMC21 was first translated from English into Norwegian according to the recommendations of the instrument's governing body. The process was documented and the quality of the translation qualitatively assessed through translator feedback, content validity exploration, and patient feedback and acceptance. The validity, reliability, responsiveness, and equivalence of the translated questionnaire were then quantitatively assessed using Pearson's Product Movement of Correlation, Cronbach's alpha, and floor and ceiling effects.

RESULTS: The EORTC-QLQ-LMC21 had good patient acceptance and performed fair to good on tests of validity, reliability, responsiveness and equivalence. The psychometric performance of the abdominal pain scale was poor due to one particular item, but there is nonetheless preliminary evidence for an acceptable level of quality and ability to meaningfully measure the health-related quality-of-life of this patient group.

Acknowledgements

This study has been made possible thanks to the help of many kind and talented people. I would like to start by thanking my supervisor, Professor Eline Aas at the Department of Health Economics, Policy, and Management at the University of Oslo, whose expertise, understanding, and patience helped guide me to the finish line. I would also like to thank Gudrun Bjørnelv, Ph.D. candidate at the Department of Health Economics, Policy, and Management at the University of Oslo, for her generosity of spirit and valuable inputs and insights.

The team at the Intervention Center at Rikshospitalet deserves a very special thank you: Resident surgeon and Ph.D. research fellow Åsmund Fretland and Special Advisor Milena Lewandowska. They took me in, showed me the ropes, and introduced me to this project that has helped me grow both personally and professionally. I will forever be grateful for their support, guidance, kindness, and faith in me.

I would also like to express my gratitude to the support staff at the gastro-surgery department at Rikshospitalet who coordinated the questionnaire dispersal to patients. Thanks also to the CoMet patients who spent their valuable time to complete the many questionnaires for this research.

And finally, I would like to thank my friends and family, especially my parents Elizabeth Hess, and Jackie and Peter Holmes, and my husband Frode Danielsen. Their unwavering support over the last two years has allowed me to complete this journey. Big thanks also to my best friend, Erica Rourke, for the hours of proofreading.

Bethany Kirsten Danielsen

Oslo, May 2015

Table of Contents

List of Figures

List of Tables

Abbreviations

HRQoL	Health-related quality-of-life
RCT	Randomized clinical trial
CRC	Colorectal cancer
QLQ-LMC21	EORTC-QLQ-LMC21
SF-36	Short-Form 36
QLQ-C30	EORTC-QLQ-C30
WHO	World Health Organization
QoL	Quality-of-life
PRO	Patient-reported-outcome
MOS	Medical Outcome Survey
QALY	Quality-adjusted-life-year
VAS	Visual analog scale
SG	Standard gamble
TTO	Time-trade-off
EORTC	European Organisation for Research and Treatment of Cancer
QLQ	Quality of life questionnaire
FIV	First intermediary version
MTMI	Multitrait multi-item
MTMM	Multitrait multi-method
MCID	Minimal clinically important differences
ES	Effect size
SRM	Standardized response mean
RS	Responsiveness statistic
FW1	First forward translator
FW2	Second forward translator
BW1	First backward translator
BW2	Second backward translator
PPMCC	Pearson's Product Movement of Correlation Coefficient

1 Introduction

Measurement is the assigning of numbers to observations in order to quantify phenomena (Kimberlin & Winterstein, 2008). In health care, this may mean measuring biological indicators for the presence of disease, or measuring a more abstract concept such as health-related quality-of-life (HRQoL). The goal of HRQoL measurement is to assess patient health as it is affected by intervention or disease, but in a way that ensures that data is free from measurement error and can be meaningfully interpreted. To accomplish this, reliable, valid, and responsive measurement instruments are needed.

HRQoL has gained increasing importance as a health outcome measure in economic evaluations performed alongside randomized clinical trials (RCTs). Economic evaluations seek to systematically measure and value the costs and benefits of two alternative interventions so they may be meaningfully compared and the best course of action identified (Drummond, 2005). Before the introduction of the patient perspective in the form of HRQoL, RCT investigators relied only on the measurement of biological indicators, such as survival time. While survival time has continued to be a very important end point measured in RCTs, the introduction of the patient perspective in health outcome measurement has provided a way for doctors and researchers to more accurately assess the actual relative benefit of treatment to the patient.

This has become especially important in the evaluation of interventions for patients with chronic and severe conditions, such as cancer. Cancer patients are some of the most severely affected by interventions that aim to either lengthen survival time or offer a cure for the disease. Because patients with cancer have many symptoms and losses of function that cannot be measured with laboratory tests, multi-dimensional health outcome measurement in the form of HRQoL is increasingly used to evaluate the effect of cancer interventions.

This study will focus on the measurement of HRQoL for colorectal cancer (CRC) patients with liver metastases, as the instrument that is the focus of this study, the EORTC-QLQ-LMC21 (QLQ-LMC21), is an HRQoL instrument designed specifically for this patient group. Using a cross-sectional study design, this analysis ultimately seeks to explore the methods used in the translation and psychometric assessment of HRQoL instruments and preliminarily assess the quality of the Norwegian QLQ-LMC21 in terms of equivalence, validity, reliability and responsiveness. The analysis focuses on the QLQ-LMC21 questionnaire and its validity, reliability, and responsiveness by exploring:

A. Content validity, patient acceptance, and equivalence achieved through the translation process
B. Psychometric validity at item-level by evaluating internal consistency, convergent validity and discriminant validity
C. Reliability, validity and responsiveness at scale-level by comparing the QLQ-LMC21 to the SF-36 and QLQ-C30
    a. The hypothesis/assumption that the QLQ-LMC21 is more sensitive (responsive) to small, yet clinically important changes in the health of patients with CRC liver metastases

This introduction included a brief introduction of HRQoL measurement in economic evaluations alongside RCTs that seek to evaluate interventions for cancer patients. In chapter two I will place this study in the context of the Norwegian setting by discussing CRC in Norway and the RCT being conducted in Oslo that seeks to evaluate the effectiveness of a new treatment method for patients with CRC liver metastasis, the CoMet study. In chapter three I will define and discuss the concepts of health, HRQoL, and health measurement. I will also discuss the development of HRQoL instruments, including the three instruments used in this study. In chapter four, I will discuss the concepts that underlie the methods for instrument translation and the development and assessment of valid, reliable and responsive HRQoL instruments that are capable of yielding meaningful data. Chapter five contains patient, data and study methods used in the analysis, followed by results in chapter six. Chapter seven will contain a study discussion, followed by a conclusion in chapter eight.

Because this study explores two processes (translation and psychometric assessment) that are approached using different methods (qualitative and quantitative, respectively), chapters four, five, and six have been divided accordingly: (1) the translation process and (2) the psychometric assessment.

2 Background

Cancer is currently the second leading cause of death in the world, including Norway. In 2012, 10,800 of the 41,900 deaths (25.8%) in Norway were attributed to cancer (Borgan, 2013). The incidence of CRC is increasing and is now the third most prevalent form of cancer and the fourth leading cause of cancer deaths worldwide, with an estimated 1.2 million cases and .6 million deaths annually (von Karsa et al., 2013). Estimates by the International Agency for Research on Cancer place CRC as the most common cancer in Europe, with 432,000 new reported cases for men and women in 2008 (Ferlay, 2010). CRC is the second most commonly diagnosed cancer in Norway, and the incidence is rising for both men and women (Hviding, Juvet, Vines, & Fretheim, 2008). Patients with CRC may experience metastasis, or spreading, of the cancer to other organs. Commonly with CRC, the metastasis may occur in the liver. In Norway, about 30% of patients present with metastases at the time of diagnosis, while another 20% develop metastases during the course of the disease (Hviding et al., 2008). Though chemotherapy may be used to manage advanced disease, hepatic (liver) resection is the only potentially curable treatment and is now offered to 20-25% of patients with liver metastases. Five-year survival rates for this surgery are currently between 30% and 58% (Abdalla et al., 2004).

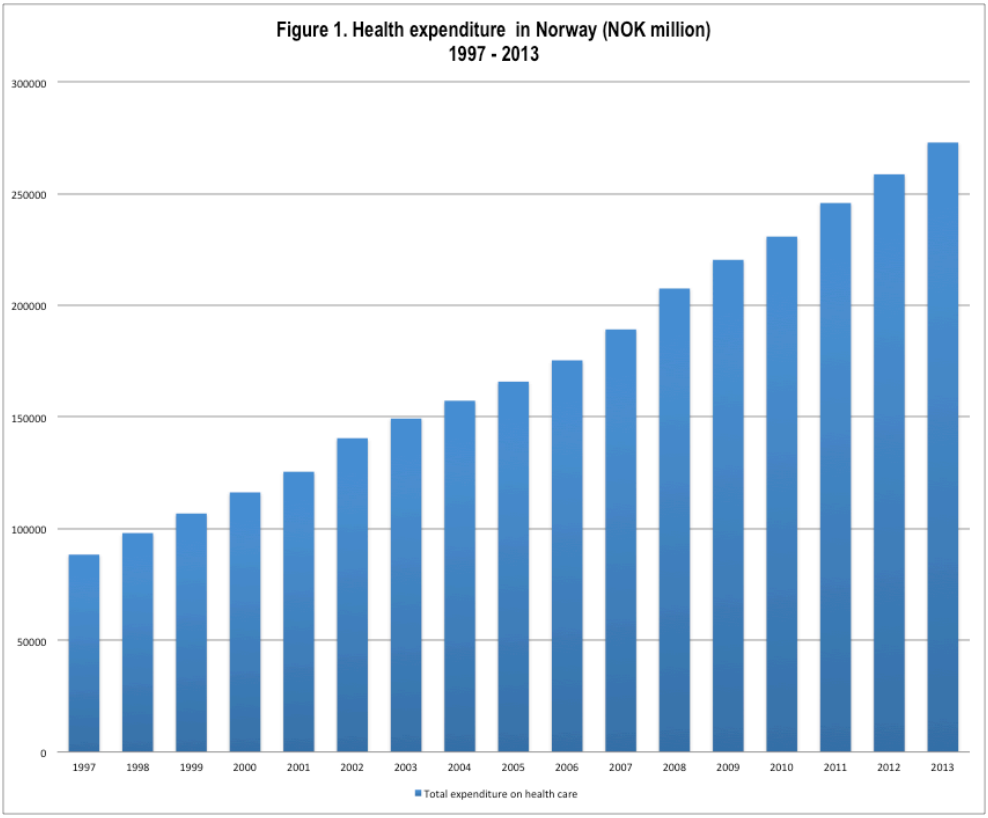
The HRQoL instrument that is the focus of this analysis, the QLQ-LMC21, was translated into Norwegian during this study for eventual use in the Oslo CoMet study. The QLQ-LMC21 is specifically designed to measure the HRQoL in patients whose CRC has metastasized to the liver, and the CoMet study is a currently operating RCT that is designed to determine whether laparoscopic liver resection of colorectal liver metastases leads to less postoperative morbidity and mortality than open liver resection. Secondary end points of the RCT include 5-year survival, disease-free and recurrence-free survival, recurrence pattern, and management of recurrence (Fretland et al., 2015).

An economic evaluation is also being conducted alongside the RCT to ascertain the hospital and societal costs and the benefits to patients as a result of treatment. Cost data will be assessed using registry data and patient questionnaires. HRQoL is currently assessed using the Short-Form-36 (SF-36). The SF-36 is given to patients before surgery (baseline), and at 1-month and 4-months post-surgery. The SF-36 is a generic HRQoL measure that is designed to measure the HRQoL in a broad range of patients regardless of the type of disease. Additionally, a subset of patients will receive the disease specific HRQoL instrument, the

QLQ-LMC21. As this measure is newly translated into Norwegian as a result of this study, it has never before been used in a RCT in Norway. Additionally, it has never before been used to evaluate the HRQoL of patients undergoing liver resection (Fretland et al., 2015).

During the last two decades, economic evaluations have been used in response to the dramatic increase in health care expenditure caused by rapidly expanding medical technology and an increase in patients living longer with more chronic diseases, such as cancer. Since 1997 in Norway, for example, total health care expenditure has increased 67.6%, from NOK 88,369,000,000 in 1997 to NOK 272,911,000,000 in 2013. Current preliminary estimates show 2014 expenditures at NOK 290,000,000,000 (Øynes, 2015). Figure 1 shows the steady increase in health expenditure in Norway since 1997.

Economic evaluations are used in clinical trials to collect data on the costs and effects of interventions (Glick, 2015). The cost of the intervention is compared to the effect data, often measured in HRQoL, to assess the relative benefit of the intervention to the patient. This assessment is important because many costly new interventions may either yield very little actual benefit to patients in HRQoL or survival time, or they may in fact detrimentally affect patient health. Economic evaluations seek to assess the relative benefit of the intervention to the patient so that policy makers can make informed decisions regarding the allocation of increasingly constrained resources in the health sector.



Source: Statistics Norway

3 Measuring health

The WHO defines health as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" (WHO, 1948). While this definition is thought by some to be idealistic and lacking an operational definition, it nonetheless recognizes the multi-dimensionality of health and has allowed for a paradigm in modern medicine that incorporates both tangible biological indicators and intangible quality-of-life perceptions.

Like the concept of health, HRQoL is multi-dimensional and can be defined in many ways, but at a minimum it involves an assessment of the physical, mental, and social effects of disease or treatment on the patient (Ferrans, Zerwic, Wilbur, & Larson, 2005). In other words, it is the way in which health is affected by disease and treatment. HRQoL is differentiated from general quality-of-life (QoL) in that it is only concerned with the ways in which disease and interventions affect health. Originally, work in this area was termed "health status" or "outcomes assessment" and could be performed in either patients or the general public. Eventually, the outcomes of these assessments being performed on patients were termed HRQoL assessment to distinguish it from general QoL, because QoL can be influenced by factors that lie outside of the health domain, such as income or environmental factors (Osoba, 2011).

The purpose of measuring health is to quantify the degree to which disease or treatment impacts the patient (ISOQOL, 2015). This purpose has gained even more importance since the 1990's with the development and expansion of the evidence-based decision-making paradigm (Bensing, 2000). The goal of evidence-based decision-making is to systematically review, appraise, and use clinical research findings to aid in the delivery of optimum clinical care to patients (Rosenberg, 1995). Increasingly, these methods are being employed to aid in decisions regarding resource allocation, namely in the form of economic evaluations alongside clinical trials.

In this chapter I will start by generally discussing HRQoL instruments and their general purpose. I will then discuss generic, utility-based, and disease-specific instruments by exploring their construction and purpose. I will also introduce and discuss in detail the measures that are the subjects of this study, the SF-36, QLQ-C30, and QLQ-LMC21.

3.1 Health-related quality-of-life instruments

Multi-dimensional definitions of health recognize that health is a product of the tangible and intangible, the objective and the subjective. This has led to two broad categories of health

measurement: objective health measurement and perceived health measurement. Objective health lies outside of the perceptions, feelings, and thoughts of the patient and can be measured with clinical indicators, such as by tumor size, weight loss, or survival time. Though objective health measurement continues to be an important component of health outcomes research, this study will focus on perceived health and its measurement. Perceived health is the aggregate subjective experience of biological function, symptoms, and functional status and can be measured using patient-reported-outcome (PROs) assessments. PROs give a subjective view of the health of patients as they experience it, without being interpreted by a clinician and recognize that patient perceptions are influenced by individual and environmental factors that vary from patient to patient (Wilson, 1995). For example, though two patients may be afflicted with the same type of cancer with identical tumors in identical locations in the body, due to biological, psychological or socio-economic qualities, and traits unique to those patients, they may experience the effects of their disease and treatments differently. PRO assessments can assess a wide variety of patient experiences, from satisfaction with treatment to the burdens of disease symptoms on day-to-day life.

HRQoL instruments are a specialized type of PRO assessment used by clinicians, researchers, and policy makers that seek to measure QoL as it is affected by disease and treatment (Blazeby et al., 2006). The development of valid, reliable, and responsive instruments since the 1970's has resulted in the assessment of HRQoL in tens of thousands of cancer patients in thousands of clinical trials. These measures have become especially useful in the assessment of the impacts of toxic and invasive treatments, such as chemotherapy and surgery, by assessing the subjective relative benefit of these treatments on patient QoL.

There are three broad types of HRQoL measures: generic, utility-based, and disease-specific. The choice of instrument largely depends on two factors: the extent to which the investigator wishes to capture health status change over time, and the desire to measure within-subject change versus between-subject change (Patrick, 1989). Though the focus of this paper is a disease-specific instrument, the QLQ-LMC21, I will begin my discussion of HRQoL instruments with generic instruments, as they are the broadest type of HRQoL instrument. Discussing them first will lay the groundwork for the increasingly narrow scope of the other two instrument types, utility-based and disease-specific, respectively.

3.1.1 Generic HRQoL instruments

Generic instruments aim for a broad assessment of HRQoL and can be administered to patients regardless of impairment, illness, or disease because the outcome is expressed in a

standard unit of measure. They are used for their ability to capture a comprehensive picture of HRQoL across all patient populations that can then be used to evaluate treatments, allocate resources, or compare disease burden between patient groups. The same generic measure can, for example, be administered to a patient diagnosed with arthritis or a patient diagnosed with CRC, and their scores can be meaningfully compared because of the standard unit of measure.

Due to their robustness and wide breadth of health states they are able to capture, generic instruments have a good capacity to measure HRQoL in a diverse set of patients. They are attractive to researchers and policy makers because they make comparisons between patient populations possible (G. Guyatt, Feeny, D., Patrick, D., 1993). When assessing the benefits of an intervention, the policy maker may be more interested in between-subject change (examining differences between individuals) at one particular time point. They may also wish to compare health outcomes across patient groups and interventions, which is not possible with disease-specific measures. They, therefore, may be more inclined to choose a generic HRQoL instrument. Examples of generic HRQoL instruments are the SF-36, the Sickness Impact Profile, and the Nottingham Health Profile.

Paradoxically, the disadvantages of generic measures are a direct consequence of their robustness. Because generic measures tend to be necessarily long in order to measure HRQoL in such a large range of patients, patients may be less likely to complete these longer measures or more likely to fill them out incompletely, leading to a lower response rate or gaps in individual data. While some measures, like the SF-36, have algorithms that try to estimate missing values based on other completed answers, some investigators may not find this an ideal solution.

Generic measures have also been found to be less sensitive, or responsive, to small yet clinically significant changes in health (G. Guyatt, Feeny, D., Patrick, D., 1993). Their ability to measure the health states of so many types of patients causes them to be less able to focus on the problems of any one particular patient group. These measures are also less able to measure small but meaningful changes in patient health over time, making them less attractive to researchers who are interested in evaluating the effects of a specific intervention on patient outcomes.

3.1.1.1 SF-36

The SF-36 was constructed in order to make comparisons of HRQoL, relative disease burden, and relative benefits of treatment between groups possible for the researchers and policy

makers involved in the Medical Outcome Survey (MOS), which was a 2-year observational study designed to help understand how specific components of the public health care system in the U.S. affect health outcomes (Stewart, 1989).

Prior to its development in 1988, there was a lack of measurement tools suitable for large-scale use across diverse patient populations. Standardized general health measures had been found useful for smaller scale research because they assessed basic human values such as functioning and emotional well-being. But due to their length, they were found to be impractical for large-scale use, such as in the MOS (J. E. Ware, Sherbourne, C., 1992). Brevity, reliability, and validity were the goals of the SF-36 developers. Its 8 dimensions and 36 items represent the most frequently measured HRQoL concepts found in widely used health surveys used since the 1970's and 1980's (J. E. Ware, Gandek, B., 1998). Today, the SF-36 is one of the most used HRQoL instruments worldwide; a literature search found over 3,000 studies that have been undertaken using the instrument.

The SF-36 currently in use is the second version of the instrument. It consists entirely of functional scales (scales intended to measure the extent to which the patient experiences various functional limitations as a result of treatment or disease) and one single item question regarding health transition. Because it is not designed to assess the symptoms associated with any one disease, it does not contain either symptom scales or symptom single-items as many disease-specific measures do.

Its 36 questions are spread over eight dimensions: physical functioning, physical role functioning, emotional role functioning, bodily pain, vitality, social functioning, mental health and general health, plus one single item for health transition status. (Kuenstner, 2002) These eight domains aggregate to form physical and mental health summary scores. The single item on health transition status is not used to calculate the scale scores, but has been found to be useful in estimating average change in health in the year previous to instrument administration (J. E. Ware, Gandek, B., 1998). In addition, the SF-36 has a utility index that uses an algorithm to derive utility scores that can be used to calculate quality-adjusted-life-years (QALYs) for use in economic evaluations.

The SF-36 is designed to be self-administered. Items are answered in a Likert scale continuum format. However, both the range of responses and the severity continuum order (not affected to very affected vs. very affected vs. not affected) for the responses are different for each question. For example, question 7 and 8 are both items in the bodily pain scale, however

question 7 has a scale with six possible answers, while question 8 only has five possible answers. Their severity continuum is, however, the same; an answer of 1 indicates no problem with pain, while the opposite end of the scale indicates a great deal of pain. SF-36 items and scales can be found in Table 1.

3.1.2 Utility-based instruments

Utility-based instruments, such as the SF-6D and EQ-5D, are a specialized type of generic instrument that measure the utility, or preference, that a patient has for a particular health state. Like generic measures, they can be given to patients regardless of diagnosis and can be used to compare outcomes across patient groups.

They often measure similar dimensions of health as generic measures, but they incorporate preference weights to calculate a single preference-based index score of health (Patrick, 1989). Preference weights are created using econometrically derived (using an estimator to represent and predict a statistical relationship) valuation methods using general population values. Subjects, usually members of the general public, are asked to imagine being in particular health states and then must score their preference for being in that health state. Valuation methods include the visual analog scale (VAS), standard gamble (SG), and time-trade-off (TTO). In the VAS method, subjects simply rate the health state on a scale from most to least preferable. The SG and TTO methods involve the subject having to value an imaginary health state by either trading life-years or risking immediate death in order to avoid the health state in question. The preference weights, also called tariffs, derived from these methods are then applied to the scores of the utility-based instruments in the calculation of the index score for a patient.

The index score derived from these instruments is scaled in reference to the absolute reference points 0, indicating death, and 1, indicating best health possible. Negative values are also possible with some utility-based instruments, and indicate a state that is experienced as "worse than death". For example, a patient with an index score of 1 is regarded as being in perfect health, while a patient who has an index score of -0.2 is considered to be in a very poor health state that is perceived to be worse than death. The index score can either stand as an overall measure of preference-based HRQoL or can be combined with life years to calculate QALYs. These single index instruments can have a considerable advantage over profile-based instruments, such as the SF-36, because of the high degree of interpretability that the index score offers.

Utility-based instruments are used in the economic evaluations of health interventions because they offer a way for decision makers to systematically compare relative disease burden and intervention effectiveness between different patient groups, as well as their ability to help generate QALYs. QALYs are used in economic evaluation to compare the cost per quality-adjusted-life-year gained from different health interventions across patient groups (Patrick, 1989).

Though desirable for their ability to standardize and compare the health benefits of interventions across patient populations and programs, these measures and their results are not without controversy. The valuation methods used to create preference weights have been criticized for being biased and not representative of the true patient experience because of the very cognitively difficult task of imagining health states that one has never experienced. Problems with the full health and death anchors have also been documented due to varying attitudes and perceptions around health and death, leading to potentially biased tariffs that may distort the index score (Augestad, Rand-Hendriksen, Stavem, & Kristiansen, 2013).

3.1.2.1 SF-6D

The SF-6D is a utility-based instrument that estimates preference-based index scores derived from a selection of SF-36 scores. The SF-6D is not a self-standing HRQoL instrument that is completed by respondents, but rather its score is derived from eleven items from the SF-36. To obtain a SF-6D score, an algorithm is applied to a completed SF-36 questionnaire that then yields a single preference weighted index score that ranges from .29 (worst health) to 1 (best health).

The SF-6D consists of six domains from the SF-36: physical functioning, role limitations, social functioning, pain, mental health, and vitality, with four to five levels of severity for each, giving a total of 18,000 possible health states. From these possible health states, 249 were selected and valued using the standard gamble technique from a UK population sample. An algorithm for transforming SF-36 data into a single index score was constructed using regression models to predict the single-index score of the SF-6D index items. (Mutebi, 2011).

3.1.3 Disease-specific HRQoL measures

Disease-specific instruments are narrower in their design than their generic and utility-based counterparts and are meant to measure the HRQoL related to a particular condition. They are designed to assess specific diagnostic groups or patient populations, often with the goal of measuring clinically significant changes in health that clinicians think are important. Examples of disease-specific measures are the Beck Depression Inventory, Arthritis Impact

Measurement Scale, and the QLQ-C30 with the QLQ-LMC21 subscale. Though they may share dimensions in common with generic measures, such as pain or mobility, disease-specific measures tend to have items, wording, and instructions that are tailored to the target patient population. They are written by consulting with doctors and patients to find the problems most associated with the specific diagnosis or symptom (Patrick, 1989).

Because disease-specific measures are designed to capture the problems experienced by a particular patient group, they are purported to be more sensitive to the health states that these patients experience and are able to detect small movements in health status, also known as the responsiveness of an instrument (Patrick, 1989). The responsiveness of disease-specific instruments is a main benefit of their use; though a generic and specific measure may both have a pain domain, the generic measure will not have questions that are designed to specifically capture a symptom associated with that particular disease or disease intervention, for example, abdominal pain for a patient diagnosed with CRC liver metastasis. In an RCT evaluating the effect of an intervention for this patient population, this facet of the health state would be lost with a generic measure and that facet of patient health would appear to be unaffected by the intervention.

It is common for disease-specific measures to be used by clinicians in their clinical work with patients or by investigators administering RCTs. Disease-specific measures are useful in achieving the goals intrinsic to both daily clinical work with patients and RCTs, namely the within-subject change (how much a patient changes over time) in health outcomes over a period of time in order to evaluate the effectiveness of a treatment. They help clinicians and researchers to distinguish between improved and unimproved patients, and accurately measure clinically significant changes in the health states of patients (Patrick, 1989).

Due to a lack of a standard unit of measure between disease-specific measures, they may be of less use when researchers wish to compare health outcomes across different diseases and programs, and can also not be used to calculate QALYs. Using only a disease-specific measure in an RCT may be limiting to researchers who wish to perform an economic evaluation alongside a clinical trial.

3.1.3.1 QLQ-C30

The European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life (QoL) Group was formed in 1980 in response to the need to advise EORTC on the design, implementation, and analysis of QoL studies in cancer clinical trials, and in 1986 they began

to develop an integrated measurement system for evaluating the QoL of patients participating in international cancer clinical trials (N. Aaronson, Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez., N., Filiberti, A., Osoba, D., Sullivan, M., , 1993). Because of practical constraints within clinical trials, the EORTC QoL Group sought to have a brief instrument that was still capable of capturing small yet clinically significant changes in health status. To achieve this, they adopted a modular approach to HRQoL instruments, with a core cancer measure, the QLQ-C30, which could be supplemented by diagnosis-specific modules, such as for CRC liver metastases or breast cancer (N. Aaronson, Cull, A., Kaasa, S., Sprangers, M., 1994). The EORTC Quality of Life Questionnaire (QLQ)-C30 is the most commonly used HRQoL instrument used in European cancer RCTs, and has been used in over 3,000 studies worldwide (EORTC, 2015).

In designing the QLQ-C30, the EORTC research group wished to build upon the conceptual and methodological framework for health status assessment that Ware et al. developed in their work with the SF-36 in the US (J. J. Ware, 1984) (J. J. Ware, Brook, R.H., Davies-Avery, A., 1980). They found this framework valuable, but tailored their work to cancer patients and placed signs and symptoms of cancer at the core of HRQoL measurement, followed by personal functioning, mental/emotional distress, general health perceptions, and social role functioning.

There were several cancer specific questionnaires in use in the 1980's, however none had been sufficiently validated. Before beginning the development of the QLQ-C30, the group defined several criteria for its construction: (1) the measure should be specific to cancer; (2) be designed primarily for patient self-administration; (3) be multi-dimensional and cover at least four basic QoL domains -- physical symptoms, physical and role functioning, psychological functioning, and social functioning; (4) be comprised primarily of multi-item scales; (5) be relatively brief. Additionally, the measure had to meet standards set for reliability, validity, and responsiveness, as well as be suitable for use cross-culturally while maintaining statistical validity (N. Aaronson, Cull, A., Kaasa, S., Sprangers, M., 1994).

The current QLQ-C30 is the third version of the questionnaire. It has 30-items and is composed of both multi-item scales and single items that reflect the multi-dimensionality of the HRQoL construct as it relates to the broad spectrum of cancer patients irrespective of body-site-specific diagnoses (N. Aaronson, Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez., N., Filiberti, A., Osoba, D., Sullivan, M., , 1993). The QLQ-C30 contains five functional scales that assess physical, role, cognitive, emotional and social functioning, three

symptom scales that assess fatigue, pain and nausea/vomiting, and a global HRQoL scale. It also contains several single-item symptom items for dyspnoea, insomnia, appetite loss, constipation, diarrhea, and financial difficulties which are meant to assess symptoms common to cancer patients (Kuenstner, 2002).

The QLQ-C30 is designed to be self-administered. Respondents are asked to consider their health during the last one week for 25 of the 30 questions. There is no time period specified for the remaining 7 questions, as it is implied that respondents generally consider their health. All questions, with the exception of the two global HRQoL questions, are answered in a Likert scale continuum format on a scale ranging from 1-4. An answer of 1 indicates "Ikke i det hele tatt", or they have not at all been affected by the health concern in question, an answer of 2 indicates "Litt", or they have been a little affected by the health concern in question, an answer of 3 indicates "En del", or that they have been partly affected by the health concern in question, and an answer of 4 indicates "Svært mye", or they have very much been affected by the health concern in question. The two questions in the global HRQoL scale are also answered in a Likert scale continuum format, however, the range of answers is expanded and ranges from 1-7. A patient who answers 1 on the scale indicates that their health as "Svært dårlig", or very bad, whereas an answer of 7 indicates they are "helt utmerket", or in excellent health. All domains and item numbers for the QLQ-C30 can be found in Table 2.

Table 2. QLQ-C30 scales and items

Scale	# of items	Item range	Item #
General QoL			
Global health status/QoL	2	7	29,30
Functional scale			
Physical functioning	5	4	1,2,3,4,5
Role functioning	2	4	6,7
Emotional functioning	4	4	21,22,23,24
Cognitive functioning	2	4	20,25
Social functioning	2	4	26,27
Symptom scale/Item			
Fatigue	3	4	10,12,18
Nausea and vomiting	2	4	14,15
Pain	2	4	9,19
Symptom single items			
Dyspnoea	1	4	8
Insomnia	1	4	11
Appetite loss	1	4	13
Constipation	1	4	16
Diarrhoea	1	4	17
Financial difficulties	1	4	28

Total number of items: 30

3.1.3.2 QLQ-LMC21

The QLQ-LMC21 is one of among 19 body-site specific modules developed by EORTC. It was designed specifically to measure the HRQoL of patients with CRC who have developed liver metastases. It was developed per EORTC guidelines through semi-structured interviews with patients and health care professionals at six cancer hospitals in the UK, France, and Germany in 2002 (Kavadas et al., 2003). The QLQ-LMC21 is the only instrument designed to measure HRQoL in patients with CRC liver metastases. Prior to its development, there were only instruments that were designed to assess the HRQoL of patients with CRC with no liver metastases, which concentrated on the gastrointestinal side effects of treatment and symptoms for this type of cancer, such as stomas and bowel and sexual function. Because the symptoms and side effects of the disease and treatment for CRC patients with liver metastases are different from CRC patients with no liver metastases, it was hypothesized that these instruments designed for CRC patients may be insensitive and irrelevant to patients undergoing treatment for CRC with liver metastases (Kavadas et al., 2003).

Blazeby et al. tested the reliability and validity of the English language version of the instrument in 2009 and found it to be a reliable and valid measure (Blazeby et al., 2009). The availability of studies on the QLQ-LMC21 is currently limited, especially in languages other than English. This study is the first to undertake the translation process in Norwegian, and Magaji et al. have recently translated the measure into the Maylasian Chinese and Tamil languages and are currently testing the measure for validity and reliability (Magaji et al., 2012).

The QLQ-LMC21 contains 21 items that are split into four scales assessing abdominal pain, activity/vigor, eating problems, and anxiety, and nine single-item symptom items that assess taste problems, dry mouth, sore mouth/tongue, peripheral neuropathy, jaundice, sexual function, nutritional issues, contact with friends, and talking about feelings (Rees et al., 2012). All domains and item numbers for the QLQ-LMC21 can be found in Table 3.

The QLQ-LMC21 is designed to be self-administered and given to patients as a seamless supplement to the QLQ-C30, meaning that the QLQ-C30 and QLQ-LMC21 are intended to be presented to the patient as one unit. The QLQ-LMC21 continues the item-numbering scheme of the QLQ-C30 and begins its numbering with question 31 (the final question of the QLQ-C30 is 30), giving the patient a total of 51 items. Also like the QLQ-C30, the items are answered in a Likert scale continuum format. All items range on a scale from 1-4, with 1 indicating not affected at all by symptoms and 4 indicating being effected a great deal.

Respondents are asked to consider their health during the last one week for 20 of the 21 questions. The remaining one question asks respondents to consider their health during the past four weeks.

Table 3. QLQ-LMC21 scales and items

Symptom scale/Item	# of items	Item range	Item #
Symptom scale			
Abdominal pain	3	4	39,40,42
Activity/vigor	3	4	37,43,44
Eating problems	2	4	31,32
Anxiety	4	4	47, 48, 49, 50
Symptom single items			
Taste problems	1	4	34
Dry mouth	1	4	35
Sore mouth/tongue	1	4	36
Tingling in fingers/hands	1	4	38
Jaundice	1	4	41
Sexual function	1	4	51
Nutritional issues	1	4	33
Contact with friends	1	4	45
Talking about feelings	1	4	46

Total number of items: 21

4 Concepts in HRQoL measurement

The measurement of HRQoL combines the desire for high quality instruments with the empirical rigor increasingly found in modern health cares systems. In this chapter, I will discuss the concepts that have been developed to aid in the creation of HRQoL instruments that are able to meaningfully measure HRQoL. Because the translation of the QLQ-LMC21 forms the basis of this study, I will begin by discussing concepts behind the translation of instruments and how researchers assess the quality of translations in terms of content validity and equivalence with the original questionnaire, and recommended translation methods for achieving content validity and equivalence. I will then continue by discussing the concepts that underpin the construction and psychometric assessment of HRQoL measures in terms of validity, reliability, and responsiveness.

4.1 Translation of HRQoL instruments

With few exceptions, most HRQoL measures are developed in the English language and are intended for use in English speaking countries (Guillemin, 1993). The increasing interest in measuring HRQoL world-wide, and increasing numbers of multi-national RCTs with a need to compare results across different countries, cultures, and language groups has led to the need for HRQoL measures to be translated and cross-culturally adapted. Because of language and cultural differences, simply transposing the measure from English into the target language will not necessarily yield a valid instrument that maintains equivalence with the original. For an instrument to yield meaningful results that can be compared across cultures, it must not only be translated linguistically well, it must also be culturally adapted to maintain the content validity and equivalence at a conceptual level. These considerations in the translation and adaptation process lead to confidence that the disease burden and health outcomes of interventions are being accurately measured (Beaton, 2000).

4.1.1 Content validity

Validity is the extent to which the instrument measures its intended constructs, for example anxiety or fatigue. Content validity is a type of validity that addresses how well the items in the instrument provide an adequate sample of all items that might measure the construct of interest. Because there is no statistical measure that can be applied to this assessment, it is a more subjective form of validity and is often left to the judgment of experts in the field and is often called face validity.

4.1.2 Equivalence

Equivalence is an important concept and consideration in the translation of HRQoL questionnaires. Equivalence is defined as the extent to which an instrument does what it is designed to do equally well in both the original and translated version. Though there is little consensus in the literature regarding its definition, equivalence essentially means that the scores from the groups taking the original questionnaire and translated questionnaire can be meaningfully compared (Herdman, 1998). A layered and systematic translation method is needed to achieve a translated measure that maintains cross-cultural equivalence. The translation procedure guidelines put forth by EORTC to guide the translations of their measures, such as the QLQ-C30 or QLQ-LMC21, are based on research by Brislin (Brislin, 1970) and Hambleton (Hambleton, 1993) and further developed by Beaton et al. in 2000 (Beaton, 2000). The back-translation framework of these theories aims to help maintain the conceptual, linguistic, cultural, and functional equivalence between the translated and original questionnaire (Dewolf, 2009).

Conceptual, linguistic, and cultural equivalence are each important components that support the overall equivalence of a translated questionnaire. Conceptual equivalence is achieved when the relationship to the underlying HRQoL concepts are the same in both the original and translated questionnaires. Linguistic equivalence is concerned with the transfer of meaning across languages, and similar effect on respondents in different languages. Cultural equivalence is concerned with assuring that takers of the measure in both languages are working under the same set of assumptions and expectations about the assessment. Some problems that may arise with cultural equivalence are differing levels of test motivation, unfamiliar test formats, and variable experiences and values (Hambleton, 1993).

4.1.3 Translation methods to achieve content validity and equivalence

Brislin, Hambleton, and Beaton recommend the following rigorous, iterative, and multi-layered back-translation process that supports the creation of an equivalent instrument. This process begins with the selection of two translators who should independently complete the forward translation from the original (source) language to the target language, and two translators who should independently complete the backwards translation from the target language back to the source language. The translators should have expertise in both the target and source languages and familiarity with both cultures, otherwise they may not be as sensitive to the unique patterns of the target and source language and culture that will allow for both a natural sounding and equivalent instrument (Hambleton, 1993). After the forward

translations are independently completed, the translators should meet to discuss their translations and come to a consensus on a synthesized translation that will be used in the backward translation process.

During the backward translation process, two translators independently translate the instrument back to the source language. This is done as a general check of the quality and content validity of the forward translation. It helps to highlight gross inconsistencies or conceptual errors, and helps to ensure that the questions have been translated in such a way that the instrument retains equivalent meaning to the source questionnaire. Backward translators can also help to fix grammatical or spelling errors in the forward translation.

Expert panels should review all translations after they are completed to further ensure the content validity and quality of the translated instrument. The role of the expert committee is to review and consolidate all versions of the questionnaire into what is considered a first intermediary version (FIV) for use in pilot testing. During the review and consolidation phase, the committee should critically evaluate the conceptual, linguistic, and cultural equivalence between the original and translated instrument to reach a consensus about any discrepancies and recommend alterations before pilot testing begins (Beaton, 2000).

The pilot test should be conducted on members of the target population to provide insight into the content validity of the instrument, as well as to identify difficult items or wording. Both the meaning of items and responses are explored to ensure that the equivalence of the measure is retained not only in theory, but also in an applied setting (Beaton, 2000).

The final stage in the translation and adaptation process should be a written technical report submitted to the instrument's developers. It is used to document the integrity of the process, the evolution of the translated instrument, and to maintain transparency of methods and quality. Developers use this report to ensure that all stages of the translation were well executed to produce a reasonable and quality translation.

Though the concepts of equivalence are evaluated by mostly qualitative methods through the iterative backward-translation process and exploration of content validity, equivalence can and should also be assessed quantitatively, and can be done by testing the retention of the psychometric properties of the questionnaire (Beaton, 2000).

4.2 Construction and psychometric assessment of HRQoL instruments

Measurement of HRQoL involves the operationalization of theoretical constructs, such as emotional role functioning or pain, and the development of instruments that are able to quantify them (Kimberlin & Winterstein, 2008). All measures are constructed by first defining several domains to measure the desired HRQoL concepts relevant to the intended patient group[1]. The range of domains that may be present in HRQoL measures is quite diverse, but instruments usually include physical, emotional (or psychological), and social domains. They may additionally include other domains such as cognitive functioning, sexuality, and spirituality (Osoba, 2011).

Domains are uni-dimensional, meaning they are intended to measure only a single concept, and can be defined as either a symptom scale or a functional scale, such as pain or cognitive function, respectively. A symptom scale is intended to measure the extent to which the patient experiences various symptoms as a result of treatment or disease, such as pain, while a functional scale is intended to measure the extent to which the patient experiences various functional limitations as a result of treatment or disease. Often times, the scoring for symptom scales and functional scales is reversed in order for the scores to be intuitively interpreted. For example, a low score on a symptomatic pain scale would indicate low symptoms of pain, while a high score on a cognitive functioning scale would indicate high cognitive functioning.

After the desired domains are defined depending on what has been deemed relevant to the patient group in question, items (questions) are written within each domain with the assumption that they will measure the underlying HRQoL domain concepts. Some measures may consist of a single global domain of general QoL, and may only ask a single question, such as "How is the quality of your life?" But because the information gathered from such a measure fails to address the multiplicity of factors that coalesce to determine HRQoL, it may not be very clinically useful. Most HRQoL measures are designed to include several different domains that consist of several items in an attempt to capture a robust picture of patient health (G. Guyatt, Feeny, D., Patrick, D., 1993).

Because HRQoL is a theoretical construct, it is more difficult to quantify and measure than traditional objective medical markers. Consequently, HRQoL researchers have borrowed strategies from the field of psychometrics, which is the study of the theory and technique of psychological measurement. The fields of psychometrics developed in response to the need of

[1] The terms domain, dimension, and scale are often used interchangeably and refer to a component of health that is to be measured within an HRQoL measure.

clinical and experimental psychologists to assess the extent to which questionnaires designed to measure abstract concepts such as intelligence or emotional functioning were truly measuring these constructs. Psychometrically designed HRQoL instruments measure the constructs underlying the many dimensions of HRQoL, such as vitality, pain, and role functioning, and provide a summary score relative to a minimum and maximum level of performance for each health concept (Lenert, 2000).

Psychometrics is primarily focused on the construction and refinement of valid and reliable measurement instruments because the foundation for all rigorous research designs is the use of sound measurement instruments (DeVon, 2007). It is an iterative process used to develop and refine measures that are valid, reliable, responsive and effective in research and clinical work. Reliability, validity, and responsiveness are key indicators of the quality of an instrument and as such, these properties are important to the development of instruments that yield accurate and relevant data (Kimberlin & Winterstein, 2008).

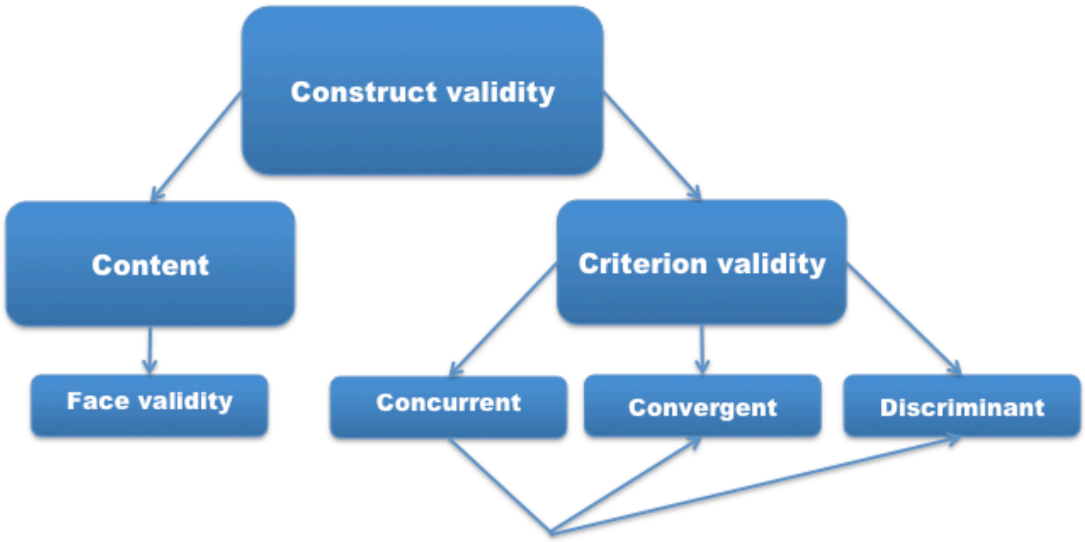
4.2.1 Validity

Validity is the extent to which an instrument measures its intended constructs. Validity requires that an instrument be reliable, but an instrument may be reliable without being valid (Kimberlin & Winterstein, 2008). In other words, though it may yield the same score for the same patient over time, it may not be measuring the constructs that it was designed to measure. For example, an instrument purporting to measure anxiety may yield the same result for the same subject at different points in time, but rather than measuring anxiety, it may instead be measuring fatigue.

Though it is common to reference the validity of an instrument, validity is actually not a property of an instrument itself. Rather, it is the extent to which interpretation of the results are warranted (Kimberlin & Winterstein, 2008). Tests of validity are intended to assess how well the instrument's results can be used to make inferences about a group of respondents. There are three main types of validity assessment: (1) content validity; (2) construct validity; and (3) criterion-related validity. A model of construct validity is shown in Figure 2. As mentioned previously, content validity is a more superficial type of validity that is judged qualitatively by considering how well the items in the instrument represent the constructs of interest. It can also be explored by assessing the number of missing responses per item, with the assumption that a question in which a high number of respondents have chosen not to answer may be upsetting or unclear in some way. Criterion-related validity involves assessing to what extent the scores of an instrument correlate with other measures of the same construct

that should be theoretically related. It can be argued that both content and criterion-related validity contribute to overall construct validity (Kimberlin & Winterstein, 2008). Content validity will be explored in this study as it relates to the quality of the translation. Criterion-related validity will be used as a proxy for construct validity of the QLQ-LMC21 and will be used to assess the extent to which the results of the QLQ-LMC21 can be used to make inferences about the HRQoL of Norwegian patients with CRC liver metastases. Support for criterion validity comes from evidence from each subtype of criterion validity: concurrent, convergent, and discriminant validity.

Figre 2. Model of validity



4.2.1.1 Convergent and discriminant validity

Convergent and discriminant validity are two subtypes of validity that make up construct validity. They are used to assess how well the instrument is measuring similar and dissimilar concepts. In other words, they are related concepts that sit on opposite sides of a spectrum. Convergent validity is the correspondence, or convergence, between constructs or items that are theoretically similar. Consequently, discriminant validity is the capability of the instrument to differentiate, or discriminate, between constructs that are theoretically different (DeVon, 2007). It is assumed that scales measure distinctly different constructs, so it is ideal that items demonstrate discriminate validity by being less correlated with other scales than with its own. For example, it is assumed and hypothesized that item 39 of the abdominal pain scale of the QLQ-LMC21 will correlate to a much lesser degree to the activity/vigor, eating problems, or anxiety scales than to the abdominal pain scale. Said in another way, item 39 should correlate to a higher degree with its own scale (the abdominal pain scale) than to other scales in the measure.

Multi-trait scaling is a way to explore whether the traits, what are being called dimensions and scales in this study, behave in the way they are expected to (Fayers, 2005). Multi-trait scaling techniques can be used to assess the convergent and discriminant validity between items and dimensions within a measure, called the multi-trait multi-item (MTMI) method, or between scales of several different instruments, called the multi-trait multi-method (MTMM). These methods are combined with a statistical test, such as Pearson's correlation coefficient, to explore the relationships between the desired items/dimensions or methods/dimensions.

4.2.1.2 Concurrent validity

Another subtype of criterion-related validity is concurrent validity. In an assessment of concurrent validity, the scores of one instrument are correlated with the scores of another instrument of high quality, called the criterion measure. Scale convergent and discriminant validity is then assessed as a way to evaluate how the instrument compares to the criterion instrument. Both instruments are concurrently administered to the same subjects at the same time point in order for the scores to be able to be meaningfully compared. Ideally, the criterion measure is the "gold standard". "Gold standard" tests are considered to be the current standard in the field and exemplify quality and correctness of results (Claasen, 2005). Unfortunately, there is currently no gold standard for HRQoL instruments, but researchers often use a well-tested and well-known measure as a substitute for a "gold standard." Apolone et al., for example, used the SF-36 as the criterion in their comparison study of the SF-36 and QLQ-C30 in their evaluation of the construct validity, of the SF-36 (Apolone, 1998).

4.2.2 Reliability

Reliability refers to the extent to which the instrument will yield the same score each time it is administered and is free from measurement error. Reliability is necessary, but it is not sufficient for a measure to be considered useful (Fayers, 2005). According to classical test theory, any score obtained by a measuring instrument consists of both "true" score, which is unknown, and "error" of the measurement (Crocker L., 1986). The true score is the score that the person would have received if the instrument were completely free of error, and the development and validation of measurement instruments, including HRQoL instruments, is in large part focused on reducing error in the measurement process (Kimberlin & Winterstein, 2008). During development of an instrument, pilot testing is often used to identify error sources so that they can be reduced or eliminated.

Reliability estimates are primarily used for three purposes: (1) to evaluate the stability of the instrument when given to the same patient at different time points (test-retest reliability); (2)

the equivalence of different observers scoring a behavior or event using the same instrument (inter-rater reliability); or (3) sets of items from the same test (internal consistency) (Kimberlin & Winterstein, 2008). Because of study design, this study will focus on internal consistency reliability estimates. However, in order to place reliability estimates within context and better describe why I have chosen to focus on internal consistency reliability, I will briefly describe test-retest and inter-rater reliability.

The test-retest reliability of the instrument is determined by administering a test to the same individual at two different time points. The strength of the correlation between the two scores is then measured. This type of reliability testing can be used for the testing of equipment, such as a scale. Ideally, the scores will be highly correlated to demonstrate a high degree of reliability and a low degree of measurement error. Test-retest reliability is not applicable in this analysis because this study is cross-sectional in design and patients were given the questionnaires at one particular point in time rather than at two or more points in time.

With inter-rater reliability, the equivalence of ratings is established when an instrument is used by different observers. These instruments are used when a researcher wishes to observe and quantify the behavior of a subject or abstract data from medical charts or when diagnoses or assessments are made for research purposes. Equivalence of ratings is established to the degree that scores from different observers of the same phenomena correlate to each other. Inter-rater reliability is not applicable to this study because there are no third party observers involved in our measurement of patient HRQoL.

4.2.2.1 Internal consistency reliability

Internal consistency estimates establishes the reliability of sets of items from the same test (Kimberlin & Winterstein, 2008). These estimates are based on the assumption that items measuring the same construct should correlate to each other. For example, items that form a pain scale should correlate highly with one another because they were placed within the pain scale with the assumption that they measure the underlying construct of pain.

Internal consistency reliability estimates can be evaluated both at the item level and at the scale level. At item level, it allows the reliability to be assessed at the micro-level. In a sense, the scale is dissected in order to view the parts that comprise the whole, and problematic items that may contribute to measurement error can be identified. This is especially helpful during measurement development or when assessing the equivalence of a newly translated measure, because problematic items that reduce instrument reliability can be identified and removed or

redeveloped. Apolone et al. focused on internal consistency reliability estimates in their comparison study of the Italian language QLQ-30 and SF-36 (Apolone, 1998). Loge et al. also chose this reliability estimate in their 1998 validation study of the Norwegian language SF-36 (Loge, 1998).

Internal consistency at scale level allows a macro-view of the scales and is usually assessed using a statistical coefficient called Cronbach's alpha. At this level the scales can be assessed for overall reliability and compared with scales that are hypothesized to measure the same construct. This method is especially useful in the assessment of HRQoL instruments that take a modular approach, such as the EORTC line of instruments, because they are developed with the assumption that the sub-modules add valuable and complimentary HRQoL to the core module. The reliability of the corresponding scales in the core measure and the sub-module can be compared to assess the quality of the scales alone and together, and whether or not the sub-module adds any value to the core measure. Bergman et al. used internal consistency estimates to evaluate and compare the reliability of corresponding scales of the QLQ-C30 and lung cancer module QLQ-LC13 to assess if the QLQ-LC13 helped to increase reliability of the complimentary scales of the QLQ-C30 (Bergman, 1994).

4.2.3 Responsiveness

The responsiveness of an instrument is defined as the ability of the instrument to measure small, but meaningful, underlying changes in HRQoL over time (Hays, 1993). Essentially, it is the ability of a measure to identify a patient as changed or not changed by an intervention. Guyatt et al. operationalized the concept of responsiveness as separate from validity and reliability in their 1985 study (G. Guyatt, Walter, S., Norman, G., 1985). As many RCTs are designed to collect and analyze data over two or more points in time, Guyatt et al. argue that the usefulness of an instrument to measure change in persons over time is not only dependent on validity and reliability, but also on the ability and sensitivity to measure minimal clinically important differences (MCID). MCID can be defined as the smallest difference in score in the domain of interest that patients perceive as beneficial and which would necessitate a change in the patient's management (G. Guyatt, Jaeschke, R., Singer, J., 1989). An operationalization of MCID has evolved to help establish more rigorous standards of interpretation for HRQoL instruments, as the correct interpretation of HRQoL scores is integral to the correct assessment of intervention efficacy. However, there is no "gold standard" for MCID, and all estimates of MCID require a study-specific value judgment (Terwee, 2003). MCID is a very

important concept in the responsiveness of HRQoL measures used in RCTs, however, it is outside the scope of this analysis due to study design and will not be a topic of focus.

There are several methods to investigate responsiveness, including evaluating effect size (ES), standardized response mean (SRM), the responsiveness statistic (RS), and floor and ceiling effects. ES, SRM, and RS are used to statistically calculate the responsiveness of an instrument when longitudinal data is available. Responsiveness by way of floor and ceiling effects can be explored visually and using response pattern distribution when cross-sectional data is available. Due to the cross-sectional design of this study, this analysis will focus on the floor and ceiling effects method. I will, however, briefly discuss the other methods that are used when longitudinal data is available.

For the longitudinal methods, the numerator is the mean change and the denominators are the standard deviation at baseline (ES), the standard deviation of change for the sample (SRM), and the standard deviation of change in response to the intervention (RS). Each has their limitations, however. The ES statistic ignores variation in change entirely, the SRM ignores information about variation in scores for clinically stable respondents, and the RS ignores information in scores for clinically unstable respondents (Fayers, 2005). All three methods, however, may be used together to gather robust information about instrument responsiveness. Additionally, when the results of a clinical trial comparing an intervention of known efficacy with a control group are available, a useful measure of responsiveness is a between group t-statistic for change scores (Fayers, 2005). The ability of an instrument to discriminate between two groups of patients adds powerful evidence for its usefulness.

4.2.3.1 Floor and ceiling effects

Responsiveness is how sensitive the instrument is to measuring particular health states. Because responsiveness is defined as how sensitive an instrument is to detecting underlying change, responsiveness hinges on the ability of an instrument to accurately capture any particular health state at any given point in time. Exploring floor and ceiling effects is a way to explore the responsiveness of a measure with a cross-section of patients, rather that using longitudinal data to compare HRQoL over several time-points.

Floor and ceiling effects are studied to assess how well an instrument is able to measure the health states of patients in relatively good health and those in poor health, respectively. Ceiling effects are the insensitivity of an instrument to measure changes in low levels of disability. For example, a measure exhibiting ceiling effects would be insensitive to the

HRQoL gains of patients in relatively good health. Conversely, floor effects are the inability of an instrument to capture HRQoL movement when patients have moderate to severe health burdens. In other words, patients are worse off than the instrument can accurately capture (Feeny, 2013). If many patients score at either extreme of a scale, the instrument will have limited ability to register deterioration or improvement, respectively (Brazier, 1999). These underestimations of the magnitude of change can bias results of the intervention and economic evaluation.

The potential for floor and ceiling effects can be assessed by analyzing response patterns. Loge et al., for example, used response patterns to analyze floor and ceiling effects in their exploration of the newly translated Norwegian language SF-36 study (Loge, 1998). Spikes at the highest and lowest response options are seen as evidence for ceiling and floor effects, respectively (Feeny, 2013). Intuitively, if a spike is seen at the high or low end of the distribution, one might infer that the instrument may have trouble differentiating between gradients in patients in either very good or very poor health.

5 Methods

This study seeks to explore the methods used in (1) translating a HRQoL instrument and (2) assessing its psychometric quality in terms of validity, reliability, responsiveness, and equivalance. Because the translation process is inherently qualitative in nature and the subsequent psychometric assessment is quantitative, the methods have been divided accordingly: (1) the translation process (qualitative); and (2) the psychometric assessment (quantitative).

The psychometric assessment has been further divided into two sections: (1) content validity, psychometric validity, and equivalence of the QLQ-LMC21; and (2) scale reliability, validity and responsiveness in comparison to the SF-36 and QLQ-C30. Content validity, psychometric validity and equivalence are tested together because they use item-level tests to assess quality. It is important to first analyze an instrument from the item level because item-level quality is the foundation upon which scale level quality rests. Scale reliability, validity and responsiveness have been given their own section for two reasons: (1) they use scale-level tests to assess quality and (2) these tests involve comparison with other instruments (the QLQ-C30 and SF-36).

5.3 Translation process methods

This section will begin with the patients, data methods, and translators used during the translation of the QLQ-LMC21 and continue with the methods used, up to and including final acceptance of the translated questionnaire by the EORTC QoL Group. Content validity methods are discussed at the end of the section.

5.3.1 Patients

As part of the translation process, the intermediary version of the QLQ-LMC21 was pilot tested on adult Norwegian patients diagnosed with CRC liver metastases being treated at Rikshospitalet in Oslo. Inclusion criteria consisted of having a diagnosis of CRC liver metastases.

5.3.2 Data

Data during the translation process was collected in two manners: (1) from the translation process excluding pilot testing and (2) pilot testing on patients. Data from the former was considered feedback from translators, the EORTC QoL Group, and the outside translation agency contracted by EORTC to advise on translation quality. This data helped to inform the translation process and ultimately the intermediary questionnaire that was given to patients in

the pilot test. Consequently, feedback from patients involved in the pilot test was considered separately as its own data.

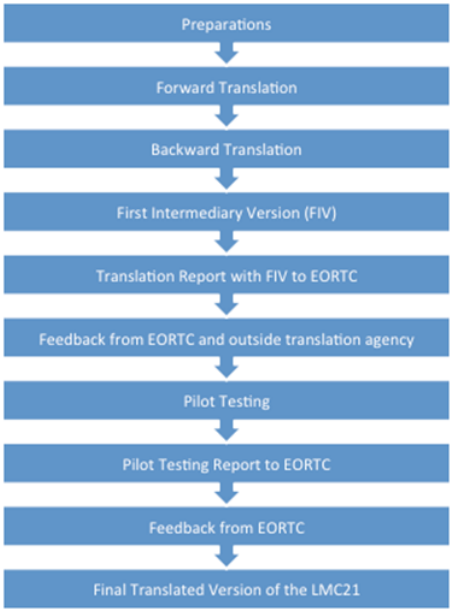
5.3.3 Translators

There were four translators involved in the Norwegian translation of the QLQ-LMC21. Two were needed to perform the forward translation from English to Norwegian, and an additional two were needed to perform the backward translation from Norwegian to English. The translators were selected based on their language skills, with all four translators having excellent spoken and written Norwegian and English skills. The two forward translators were current masters students in the Health Economics, Policy, and Management program at the University of Oslo. The first forward translator (FW1) lived in Norway as a small child and was educated in Norway through grade school. She then attended an American high school in the Dominican Republic, and afterwards returned to Norway to pursue higher education. She has resided in Norway for 16 years. The second forward translator (FW2) is a native Norwegian speaker who lived in Norway until she was 5 years old, then lived in the United States for 20 years and received formal education there through the college level. She has resided in Norway for the last 3 years. The two backward translators were previous masters students in the Health Economics, Policy, and Management program at the University of Oslo. Both the first backward translator (BW1) and the second backward translator (BW2) are native Norwegians who have lived and attended American schools both in Norway and abroad.

5.3.4 Translation process

The translation of the QLQ-LMC21 is the result of a collaborative effort between this author (acting as the translation coordinator) and the EORTC QoL Group. Translation was a multi-step process that took place over the span of 7 months. Figure 3 illustrates the process.

Figure 3. Translation process of the QLQ-LMC21



5.3.4.1 Preparation: April 28 2014

Communication with the EORTC Quality of Life Group was established during the preparation phase. During this phase they provided this author with the version of the QLQ-LMC21 that would be used in the translation process. Because EORTC maintains a large bank of questionnaire items that can be drawn upon in the creation of new EORTC instruments, some of the items in our initial version of the questionnaire had already been translated into Norwegian. The items that had been pre-translated on the initial QLQ-LMC21 are shared with other EORTC measures. The questionnaire instructions, 13 of 21 questionnaire items, and response categories were pre-translated into the Norwegian language, leaving only eight items to be translated during this process. During the preparation stage, two forward and two backward translators were recruited to perform the translations.

5.3.4.2 Forward translation: May 9 to May 16 2014

Both forward translators were sent an electronic version of the pre-translated questionnaire and given five days to independently complete the forward translations of the eight questionnaire items. They checked the instructions, pre-translated items, and response categories for errors as well. After the forward translations were completed, the translation coordinator and the two translators met via a video conference call to discuss and reconcile the two translations into a single cohesive questionnaire.

During the call, the two versions of the translation were compared and discussed by the two translators, while this author coordinated the discussion and acted as mediator. The following criteria were used for a final conclusion:

A. If both versions were identical, no changes were made.
B. If there was a difference, the most appropriate translation was chosen:
    a. The sentence as close as possible to the original meaning, but also fitting into the Norwegian cultural setting.
    b. When a sentence in the two versions had the same meaning we chose the translation that patients would be more likely to understand and use.
    c. The shortest sentences.

The following items were translated and discussed:

33. Have you worried about losing weight?
41. Have your skin or eyes been yellow (jaundiced)?
44. Have you felt lacking in energy?
46. Have you had trouble talking about your feelings to your family or friends?
47. Have you felt stressed?

48. Have you felt less able to enjoy yourself?
50. Were you worried about your family in the future?
51. Has the disease or treatment affected your sex life (for the worse)?

The two forward translations versions were reconciled into one document and ready for translation from Norwegian back to English.

5.3.4.3 Backward translation: May 26 to June 11 2014

To ensure that the eight translated items retained their original English meanings, the items were translated from Norwegian back to English (backward translation) by the two backward translators. Both translators were sent an electronic file containing only the eight Norwegian questionnaire items. To prevent the translators from searching for the original English version of the questionnaire and becoming biased in their translations, they were not informed about which questionnaire the items were from, only that they were from a HRQoL instrument for cancer patients. The translators were given 16 days to independently complete the backward translations. After the translations were completed, the translation coordinator and two translators met in-person to discuss and reconcile the two translations into one cohesive document. The original English version of the questionnaire was used during the meeting as a reference point, helping to guide the reconciliation process.

The two backward translations were reconciled into one cohesive document, as well as any recommended changes to the forward translation due to conceptual weakness or grammatical or spelling errors.

5.3.4.4 Feedback from EORTC: July 2014

After the forward and backward translations were completed, a translation report detailing the translation process and preliminary first intermediary version of the QLQ-LMC21 was prepared and sent to EORTC so they could provide feedback, recommend changes, and send the questionnaire to an outside translation agency for review.

5.3.4.5 Feedback from translation agency: August 2014

EORTC sent the questionnaire to an outside translation agency to provide expert feedback, recommend changes, and ensure that the instrument was ready for pilot testing.

5.3.4.6 Pilot testing: September 2014 to November 2014

After feedback from the translation agency was received, the questionnaire was pilot tested on ten patients diagnosed with CRC liver metastases at Rikshospitalet to gather information about content validity, patient acceptance, and any potentially problematic items.

5.3.4.7 Final Acceptance of the QLQ-LMC21 by EORTC: November 2014

Comments gathered from patients during the pilot testing were sent to the EORTC QoL Group for additional feedback and eventual final acceptance of the instrument.

5.3.5 Content validity

Content validity of the QLQ-LMC21 was explored during all levels of the translation process. Content validity was assessed during the forward and backward translations based on translator feedback, feedback from EORTC and the outside translation agency, and from patient feedback during pilot testing. During pilot testing, patients were interviewed by one of the principle investigators of the CoMet study (a Norwegian surgeon who is both fluent in Norwegian and very familiar with this patient population) and given a comment sheet after completing the questionnaire. They were asked to identify any items that they did not understand or felt were problematic in some way. They verbally communicated any feedback to the interviewer, as well as logged it on a comment sheet. Content validity was assessed qualitatively by way of patient feedback.

5.4 Psychometric assessment methods

This section begins with the patients and data used in the psychometric assessment of the QLQ-LMC21. The psychometric assessment has been divided into two sections: (1) content validity, psychometric validity, and equivalence of the QLQ-LMC21; and (2) scale reliability, validity and responsiveness in comparison to the SF-36 and QLQ-C30.

5.4.1 Patients

The QLQ-LMC21 was given to adult Norwegian patients participating in the CoMet study at Rikshospitalet in Oslo. Inclusion criteria consisted of a diagnosis of CRC liver metastases, and participation in the CoMet study. Patients included in this study were at various time-points in the study. For example, some had not yet had the trial surgery, while others were at four weeks or four months past their surgery.

5.4.2 Data

Patients were given a 10-page packet consisting of the SF-36, QLQ-C30, and QLQ-LMC21 questionnaires (in this order) during clinic visits and by post between December 2014 and April 2015. Patients self-administered the questionnaires at home and sent them back to Rikshospitalet by post. Pre-addressed and pre-postage paid envelopes were included to help facilitate a higher response rate. The questionnaires included patient identification numbers so that patient demographics could be collected and used in the analysis. Some patients

completed the questionnaire four weeks after randomization into the study, while others completed the questionnaire at four weeks or four months after surgery.

5.4.2.1 Instrument scoring

Scores for all instruments were calculated in Microsoft Excel based on the instructions of the instrument developers.

QLQ-C30 & QLQ-LMC21

The scoring methods for the QLQ-C30 and QLQ-LMC are the same because they are both developed by EORTC. First, the raw score of each questionnaire was computed by averaging the items in each scale and also the single items:

Raw Score = ***RS*** = $(I_1 + I_2 + ... + I_n) / n$, where $I_1 + I_2 + ... + I_n$ are the items in the scale

Next, the raw scores were linearly transformed to obtain a score, *S*, from 0-100:

Functional scales: S = [1-(RS-1)/range]*100

Symptom scales, symptom single items, and global health status: S = [(RS-1)/range]*100

Range was defined as the difference between the maximum possible value of RS and the minimum possible value. For example, the lowest value a patient could choose on a QLQ-LMC21 item scale was one, while four was the highest, giving a range of three.

As an example, the emotional functioning scale score of the QLQ-C30 was calculated in the following way:

RS = $(Q_{21} + Q_{22} + Q_{23} + Q_{24}) / 4$

Scale score = [1 - (RS-1) / 3]*100

All of the scales and single-item scores range from 0 to 100, but the interpretation of functional scales and symptom scales and single items are different. A high score for a functional scale represents a high level of functioning, a high score for the global health status scale represents a high QoL, but a high score for a symptom scale/item represents a high level of symptomatology. (EORTC, 2001)

SF-36

The SF-36 is scored in two steps. First, the response for each item is recoded with a value from 0-100 depending on the wording of the question (positively or negatively worded) and the range of values on the Likert scale. Values for the recoding of items can be found in Table 4. Second, an average value is calculated for each of the recoded items in each scale. Missing data was ignored and the scale score calculated without the missing item, however if more than 50% of the items of any one scale are missing, the scale score was not calculated.

Table 4. SF-36 Recode Key

Item Numbers	Original Response	Recode Score
3a, 3b, 3c, 3d, 3e, 3f, 3g, 3h, 3i, 3j	1	0
	2	50
	3	100
2, 4a, 4b, 4c, 4d, 5a, 5b, 5c, 9b, 9c, 9f, 9g, 9i, 10, 11a, 11c	1	0
	2	25
	3	50
	4	75
	5	100
7	1	100
	2	80
	3	60
	4	40
	5	20
	6	0
1, 6, 8, 9a, 9d, 9e, 9h, 11b, 11d	1	100
	2	75
	3	50
	4	25
	5	0

5.4.3 Content validity, psychometric validity, and equivalence of the QLQ-LMC21

All statistical analysis was performed using SPSS version 21. Microsoft Excel was also used for assessment of content validity, graphs and exploration of the data.

The content validity, psychometric validity and equivalence of the Norwegian language QLQ-LMC21 were explored first. These tests were conducted together because they explore the QLQ-LMC21 at (1) item level and (2) alone and not in comparison with any other measure. Content validity was evaluated by assessing the number of missing items. Establishing psychometric validity involves tests of both reliability and validity at the item level. Item internal consistency (a test of reliability) and item convergent and discriminant validity (tests of validity) were conducted using Pearson's Product Movement of Correlation (PPMCC) and

a multi-trait multi-item (MTMI) method matrix. Equivalence was assessed by patient acceptance, and tests of convergent and discriminant validity.

5.4.3.1 Content validity

Content validity of the QLQ-LMC21 was explored by assessing the distribution of responses for all items. Each item of the QLQ-LMC21 was tallied in an Excel spreadsheet in one of five categories, with four of the categories corresponding to the item range (1 – 4) and one category corresponding to a missing value. For example, if a patient answered 1 ("Ikke i det hele tatt") for question 39, that response was tallied in category 1 for item 39. If the patient did not answer question 39, a tally was placed in the "missing" category. Patterns of missing items were then assessed.

5.4.3.2 Psychometric validity

Item internal consistency, item convergent validity, and item discriminant validity were assessed using Pearson's Product Movement of Correlation in combination with the MTMI method.

Pearson Product Movement of Correlation

Pearson's coefficient, also known as the Pearson Product Movement of Correlation Coefficient (PPMCC), is often used to establish the validity of a measure. It is used as a measure of the degree of linear dependence between two variables and is used to test the correlation between sets of data as a measure of how well related they are. The formula for PPMCC can be expressed as:

$$\rho_{X,Y} = \frac{\sum[(x_i - \bar{x})(y_i - \bar{y})]}{\sigma_X \sigma_Y}$$

Where X and Y being sets of values from two different scales, x_i and y_i represents each individual value in each scale. By calculating the difference of each value from the mean of each scale and multiplying them together, we get a new set of products of which the sum represents the covariance for X and Y. Because the values of X and Y are in different units, the covariance will be in the product of units for the scales, making it difficult to reason about the significance of the covariance. By dividing the covariance with the product of standard deviations for X and Y, we get a normalized value between -1 and 1 called the Pearson's correlation coefficient.

As an example, we have the measurements of heights (cm) and weights (kg) of ten people, and the means of these heights and weights. If we wanted to find the PPMCC of these

subjects, we would subtract the mean of heights from the height of each person, multiplying each of these differences with the differences of weight from the mean for each person, and multiply the differences together. The sum of all these products would be the covariance, a measurement in cm-kg, which is difficult to interpret because cm and kg are measured on different scales. Dividing the covariance with the product of standard deviations would give a unit-less, normalized value within a continuous range between -1 and 1. Where the extreme values -1 and 1 represents a strong negative or positive correlation, respectively. And a value of 0 represents seemingly no correlation. PPMCC assumes a linear relationship between variables. A correlation of 0 might also mean the scales are not linear.

When using PPMCC to evaluate validity of scales within an HRQoL measure, items within a subscale should correlate highly with each other (above .40) to demonstrate item convergent validity. Item discriminant validity is supported if the correlation between the item in question and its own scale is significantly higher (defined as two standard errors) than the correlations between the item and other scales. Mean scores and standard deviations of items within scales should also be very similar in order to confirm scaling assumptions (J. E. Ware, Gandek, B., 1998).

Multi-trait multi-item (MTMI) method

The MTMI method allows the correlations between items and scales to be examined in order to gather evidence for convergent and discriminant validity. When an instrument is developed, items are chosen for a scale with the assumption that they are measuring the intended construct. Ideally, measures should demonstrate convergent validity (to show that the item is measuring the hypothesized concept) by having items that correlate highly to its own scale. They should also demonstrate discriminant validity (to show that the item is not measuring an unintended concept) by correlating less to other scales. Item/scale correlations are the fundamental elements of multi-trait scaling, and constitute the MTMI correlation matrix (Fayers, 2005).

Table 5 shows an example of a MTMI correlation matrix. Three different scales, or traits, are defined by items 1-10. Each row of the matrix contains correlations between the scores for one item and all hypothesized scales. Each column contains correlations between the scores for one scale and all the items in the analysis. Correlations between an item and its own scale should be corrected for overlap, so that estimates of the item-scale relationship are not erroneously inflated. Scale scores corrected for overlap are calculated by removing the item in question and computing the total scale score from the remaining items in that scale (Hobart,

Williams, Moran, & Thompson, 2002). A variable *X* is correlated with the scale score *S*, where *S* = sum (*X* + *other scale items*), because *X* is in both terms. To examine to correlation of *X* with the scale, *S* must first be recalculated without *X* (Fayers, 2005).

Item convergence is supported if an item correlates substantially (above .40) with the scale it is hypothesized to represent. In Table 5, evidence for item convergent validity is supported because all items are highly correlated (>.40) with their own scales after the scales have been corrected for overlap. Item discriminant validity is supported if the highest correlation in a row of the matrix is the correlation between the item and the trait it is hypothesized to measure, and this correlation is significantly larger (>2SE) than other correlations in the row. The SE is the standard deviation of the sampling distribution of the statistic and can be used to refer to an estimate of the standard deviation derived from a particular sample. The SE acts as a confidence interval for the matrix, and because statistical significance is usually considered to be present when a test statistic lies within two standard deviations of the mean, the same assumption is applied to the MTMI matrix. If the correlations between the items and their hypothesized scales are within two SE, it cannot be stated with confidence that the between-scale correlation is weak enough to be considered discriminant in nature. The data in Table 5 shows evidence for discriminant validity because the correlations between the items and their hypothesized scales are greater than two SE (2 x .03 = .06) above the correlations between the items and the other scales. For example, if the correlation between item 1 and scale 2 had been .74 or higher, discriminant validity would not be supported for that item.

Table 5. Example multi-trait multi-item (MMTI) correlation matrix

		Scale 1	Scale 2	Scale 3
Scale 1	Item 1	**.80**	.20	.10
	Item 2	**.78**	.15	.11
	Item 3	**.82**	.20	.10
	Item 4	**.81**	.22	.15
Scale 2	Item 5	.30	**.75**	.24
	Item 6	.28	**.85**	.22
	Item 7	.31	**.81**	.21
Scale 3	Item 8	.25	.20	**.83**
	Item 9	.18	.21	**.82**
	Item 10	.16	.20	**.82**

Underlined values corrected for overlap
Standard error of the correlation matrix = .03

Item internal consistency reliability

Item internal consistency is a test of reliability and is the degree to which the items within a scale correlate with the other items in a scale, the overall scale, and different scales in an instrument. For example, it can be used to assess how well item 39 of the QLQ-LMC21

abdominal pain scale correlates with items 40 and 42 (which are also in the abdominal pain scale), the overall abdominal pain scale score, and the activity/vigor scale, eating problems, and anxiety scales. It is evaluated to assess to what degree the items form a valid scale that measures what it purports to. Ideally, item 39 would correlate highly with items 40 and 42 (which would demonstrate item internal consistency), and correlate to a lesser degree to the other three scales (which would demonstrate discriminant validity).

To avoid inflated correlations between items and their own scales, scales were corrected for overlap by removing the item in question and computing the total scale score from the remaining items in that scale. For example, the Activity/Vigor scale consists of three items: questions 37, 43 and 44. To correct for overlap, the Activity/Vigor score scale was calculated three times, each time removing one of the questions from the scale. The scale score was calculated once without question 37 (only with the raw scores from questions 43 and 44), then without question 43 (only with the raw scores from questions 37 and 44), and finally without 44 (only with the raw scores from questions 37 and 43). The individual items from each scale were then tested against both the corrected and uncorrected scale scores. Item internal consistency was considered the high and low range of Pearson's correlation scores for an item in its own scale. For example, the item internal consistency for items 39, 40, 42 (abdominal pain scale) was considered only for the range of scores within the abdominal pain scale.

Item discriminant validity

Item discriminant validity range was considered to be the high and low range of PPMCC scores for the items and the scales for which they were not considered a part of. For example, discriminant validity for items 39, 40, and 42 (the abdominal pain scale) were observed for the activity/vigor, eating problems, and anxiety scales.

Item convergent validity test

An item convergent validity test was performed to assess the number of times the correlations of items within its own scale corrected for overlap had a higher PPMCC than .40. The number of items in each scale that scored above .40 were summed and divided by the total number of items in the scale. For example, in a scale of three items, if two of the PPMCCs of the items were above .40, then the convergent validity percentage would be 66.7%. The test can be illustrated in the following equation:

$$\text{Item convergent validity} = \frac{\sum_{r_i}^{n}[r_i > 0.4]}{\text{N}}$$

Where r_i represents each correlation, if r_i > 0.4 a 1 is tallied, otherwise a 0 is tallied. The final tally equals the number of correlations above 0.4, and dividing this by the number of correlations, N, gives the item convergent validity.

Item discriminant validity test

An item discriminant validity test was performed to assess the number of times the correlations of items within its own scale was significantly higher than its correlations with other scales. The test was considered a success when the item in the matrix row was (1) the highest correlation in the row; and (2) greater than two SE above all correlations. For example, if the SE of the correlation matrix was .03, two SE above would be .06. The number of significantly higher correlations was divided by the total number of correlations the item had with the other scales. For example, if an item in a pain scale were to have a PPMCC of .50 with the pain scale (its own scale), .50 would need to be the highest correlation in the item's row, and a correlation lower than .44 with any other scale would be considered a success of the item discriminant validity test. The total number of successes was divided by the total number of correlations. The test can be illustrated in the following equation:

$$\text{Item discriminant validity} = \frac{\sum[\Delta r > 2SE]}{\text{N}}$$

Where Δr is the single difference between two correlations (the item with its own scale subtracted from the correlation of the item with another scale), SE is the standard error of the matrix table, and N is the number of total correlations for all items within a scale to the other scales. If Δr is higher than two SE, the test is counted as a success. This test is completed for all items in a scale, and then the sum of successes is divided by the number of correlations of items with the other scales.

5.4.3.3 Equivalence

Equivalence with the English language version was established through the results of the content validity exploration and item convergent and discriminant validity tests. Equivalence was considered satisfactory when the instrument had a high degree of patient acceptance, and had a high degree of both convergent and discriminant validity (items within a scale scored above a .40 when corrected for overlap, and when items between scales scored two SE below the within-item-scale correlation).

5.4.4 Reliability, validity and responsiveness

To further assess the quality of the instrument, reliability and validity at the scale level were explored. Scale reliability of the QLQ-LMC21 was assessed using two methods: (1) using

tests of internal consistency reliability and (2) using reliability estimates of the comparable scales of the SF-36 and QLQ-C30. Scale validity was assessed using tests of concurrent validity by comparing comparable scales of the QLQ-LMC21 to the SF-36 and QLQ-C30. Responsiveness was assessed in terms of floor and ceiling effects.

Explorations of reliability, validity, and responsiveness were conducted together because (1) they examine the QLQ-LMC21 at scale level, rather than item level, and (2) they involve comparison with the SF-36 and QLQ-C30. These tests, however, are conducted using three different methods. Reliability is assessed using Cronbach's alpha because Cronbach's can only be used to assess reliability at the scale level. Concurrent validity is assessed using PPMCC, because as the goal is to explore how well comparable scales relate to each other, a measure of correlation is needed. Floor and ceiling effects are explored using response distribution between comparable scales.

Comparable scales for all comparisons were: (1) abdominal pain scale of the QLQ-LMC21, pain scale of the QLQ-C30, and bodily pain scale of the SF-36; (2) activity/vigor scale of the QLQ-LMC21, fatigue scale of the QLQ-C30, and vitality scale of the SF-36; (3) anxiety scale of the QLQ-LMC21, emotional role functioning scale of the QLQ-C30, and mental health scale of the SF-36[2]. The items from these scales can be found in Table 16 in the appendix.

5.4.4.1 Reliability

Scale reliability estimates are assessed using Cronbach's alpha. Cronbach's alpha was used in assessments of both internal consistency reliability and reliability estimates of the comparable scales of the QLQ-LMC21, SF-36, and QLQ-C30.

Cronbach's alpha

Cronbach's alpha is a function of the average inter-correlations of items and the number of items in the scale, and it will generally increase as the inter-correlations among test items increase. Unlike PPMCC, which can provide data at item level, Cronbach's alpha should only be used to yield data at scale level. Cronbach's alpha tends to increase as the number of items in a scale increase, so using Cronbach's for a single item would yield a very low and inaccurate reliability estimate (DeVon, 2007).

[2] The SF-36 also has an emotional role scale, but upon investigation, the questions from the SF-36 mental health scale appeared to be more comparable to the QLQ-C30 emotional role functioning scale, and as such, a decision was made to compare the scales that were more similar in content rather than identical in name. The emotional role scale of the SF-36 contains items that ask how the respondent feels about how much the disease has impacted their everyday activities, while the mental health scale asks questions regarded their general emotional state.

Cronbach's alpha is used to measure the extent to which the items in a scale correlate with themselves and each other, or in other words, the extent to which they measure the same or differing constructs (Loge, 1998). Cronbach's alpha coefficients range from 0 to 1, with higher coefficients indicating higher levels of reliability. Nunnally et al. have recommended that an alpha of .70 or higher is needed to establish the reliability of a scale for group comparisons, such as group comparisons made in RCTs and other clinical studies (Nunnally, 1978). However, very high reliability (above .95) can indicate that the items within the scale are redundant. (Nunnally, 1978) George and Mallery have also recommended a gradient interpretation for alpha coefficients: > .9 – Excellent, > .8 – Good, > .7 – Acceptable, > .6 – Questionable, > .5 – Poor, and < .5 – Unacceptable. (George, 2003)

Cronbach's is frequently used in tests of reliability and is a function of the number of items in a test, the average covariance between item pairs, and the variance of the total score. It is expressed in the following formula:

$$\alpha = \left(\frac{K}{K-1}\right)\left(1-\frac{\sum V_i}{V_T}\right)$$

In the formula, K represents the number of items in the scale, while V_i represents each item's individual variance in the scale, of which we divide with V_T, the total variance of the scale.

Cronbach's alpha is expected to be a value between 0 and 1, however in practice, the value can range from negative infinity to 1. Negative Cronbach alphas imply a negative average covariance among items. Negative alphas can occur in small sample sizes due to sampling error, which produces a negative average covariance.

Internal consistency reliability

Internal consistency at scale level was assessed. Between-scale and overall scale reliability of the QLQ-LMC21 was tested using overall scale scores and Cronbach's alpha. Because it is desirable that scales in a measure do not correlate highly with each other, an alpha below .70 was desirable for the between-scale tests. An alpha above .70, however, was desirable for overall scale reliability, as that is the recommended minimum to establish scale reliability.

Reliability estimates of comparable scales

The internal consistency reliability of the pain, fatigue, and mental health scales of the QLQ-C30, QLQ-LMC21, and SF-36 were compared to each other using Cronbach's alpha. The

reliabilities of these scales were first calculated separately. Because the SF-36 is separate from the QLQ instruments, the SF-36 scales remained isolated. The items from comparable scales of the QLQ-C30 and QLQ-LMC21, however, were then combined to explore whether or not the reliability of the scale was worsened or improved with the combining of the items. For example, the reliability of the pain scales for the QLQ-C30 (containing two items) and QLQ-LMC21 (containing three items) were calculated separately. The items in the scales were then combined for a total of five items, and the Cronbach's alpha calculated. If the alpha was improved by the additional items from the second scale, the scales from both measures can then be seen as playing an integral role in the reliability of the combined use of the QLQ-LMC21 and QLQ-C30. The reliability scores of the SF-36 scales were used only as the "gold standard" measure for comparison of the other two measures.

Due to reverse scoring of 6 of the 11 items from the 3 scales of the SF-36 (items 7 and 8 from the pain scales, 9a and 9e from the vitality scale, and 9d and 9h from the mental health scale), the raw scores of the positively worded questions (or questions which had reverse scoring) were transformed so that a high score indicated better health. This maintains the consistency of the overall scoring scheme of the functional scales of the SF-36, where a higher score indicates better functioning and better health. All items were answered on a 5-point Likert scale with the exception of item 7, which is answered on a 6 point Likert scale. For example, question 9a from the Vitality scale asks "…how much of the time in the last 4 weeks did you feel full of life?" The answers range from: 1 = All of the time; 2 = Most of the time; 3 = Some of the time; 4 = A little of the time; 5 = None of the time. To maintain the consistency with the negatively worded questions, these answers were transformed into: 1 = None of the time; 2 = A little of the time; 3 = Some of the time; 4 = Most of the time, 5 = All of the time.

As with internal consistency reliability, a Cronbach's alpha of .70 was considered the minimum to establish scale reliability. Scores were also interpreted based on the gradient suggestions by George and Mallery.

5.4.4.2 Validity

PPMCC for comparable scales of the QLQ-LMC21, SF-36, and QLQ-C30 were compared in a multi-trait multi-method (MTMM) matrix to investigate concurrent validity between the instruments.

Multi-trait multi-method (MTMM) matrix

The MTMM method was developed by Campbell and Fiske in 1959 as a way to simultaneously measure the correlations between different constructs that have been measured

by multiple methods (Campbell, 1959). It is one of the most common methods used to explore the construct validity of an instrument. The MTMM can be used whenever there are two or more constructs being measured with two or more methodologies. The methodologies that can be compared using the MTMM can either be completely different methodologies, like a standard paper and pencil questionnaire and direct observation by a researcher, or separate questionnaires that measure similar constructs. PPMCC is frequently used in the MTMM method as a way to measure the correlation between concepts.

A matrix of traits and methods is built and convergent and discriminant validity is assessed based on the pattern of correlations between the traits and methods. Different measures of the same construct should correlate highly with each other (convergent validity) and different constructs should show low correlation with each other (discriminant validity) within the matrix (DeVon, 2007). Figure 4 demonstrates an example of a MMTM matrix.

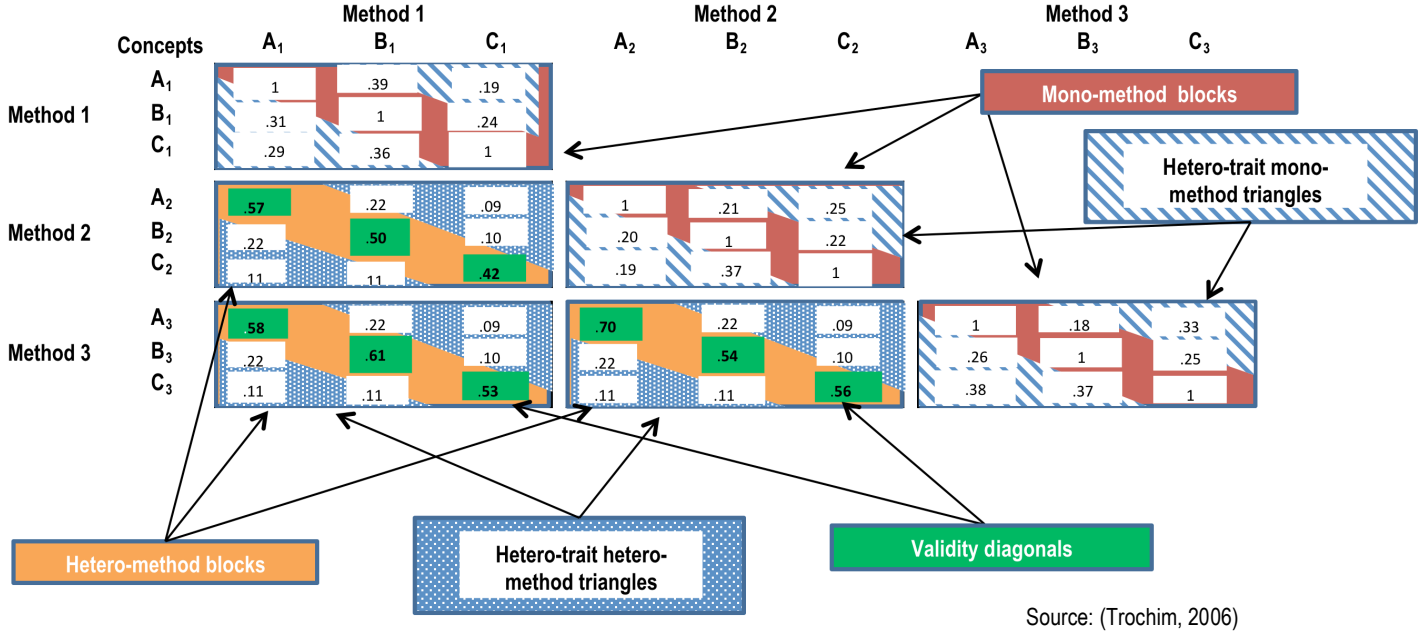
Figure 4 is arranged by method and concept. For example, methods 1 and 2 could be different HRQoL instruments, and method 3 could be direct observation scores by a researcher. The concepts are the HRQoL concepts that the researchers wish to compare between methods, such as physical functioning, cognitive functioning, and fatigue. The methods are split up into blocks: mono-method blocks, and hetero-method blocks. The mono-method blocks consist of the correlations that share the same method of measurement. Within the mono-method blocks lay the hetero-trait mono-method triangles. These triangles contain the correlations for concepts that share the same method of measurement. It is possible to achieve relatively high correlations for these items because measuring different concepts with the same instrument results in a correlated measure. The mono-method blocks also include the correlations between the same concept within the same measurement. These scores will always be one because an item will always be perfectly correlated with itself.

Conversely, the hetero-method blocks consist of the correlations that do not share the same methods. It is in the hetero-method blocks where convergent and discriminant validity can be assessed. The validity diagonals show the convergent validity, in other words, how well the same concept from different measurement methods correlate with each other. In figure 4 for example, the convergent validity of concept C1 and C3 is .53. Because these two measures are of the same concept, they are expected to highly correlate.

The hetero-trait hetero-method triangles are where the discriminant correlations are expected to be found because these correlations differ by both measurement method and concept. The

correlation, for example, between C1 and B2 is lower than the correlation between C1 and C3 functioning and fatigue at .10. This is expected because C1 and B2 are measuring different concepts, such as physical functioning.

Figure 4. Example of an Multitrait Multimethod (MTMM) matrix



Source: (Trochim, 2006)

Concurrent validity

Though there is no "gold standard" in HRQoL measures, the Norwegian language SF-36 has been used in Norway since its translation by Loge et al. in 1998 (Loge, 1998). The SF-36 will be treated as a "gold standard" for our purposes to explore the reliability and concurrent validity of the QLQ-LMC21 and test how well the scales of the QLQ-LMC21 correlate with other well-established variables that measure the same aspects of health. Additionally, complimentary scales of the QLQ-C30 are also compared to the QLQ-LMC21 and SF-36.

Concurrent validity is the most commonly used type of construct validation. It focuses on the extent of the correlation among several measure of the same concept (Apolone, 1998). Concurrent validity was tested at scale level by constructing a MTMM matrix and comparing the pain, vitality/fatigue, and mental health/anxiety scales of the QLQ-LMC21, QLQ-LMC30, and SF-36 and assessing the convergent and discriminant validity of the scales in the hetero-method blocks of the matrix. The correlations in the validity diagonals are considered to be evidence of convergent validity between instruments, while the correlations in the hetero-trait hetero-method triangles were considered to be evidence of discriminant validity between instruments. PPMCC values of .40 were needed in the validity diagonals to establish convergent validity. Correlations lower than those found in the hetero-trait hetero-method triangles were assumed to establish discriminant validity.

5.4.4.3 Responsiveness

Responsiveness of the QLQ-LMC21 in terms of floor and ceiling effects was explored using Excel to plot response distributions.

Floor and ceiling effects

Floor and ceiling effects were first tested by looking at the mean, standard error, worst possible score percentage, and best possible score percentage for the pain, vigor/fatigue, and mental health/anxiety scales for the QLQ-LMC21, QLQ-C30, and SF-36. Scale scores were examined for each patient for the presence of the best and worst possible scores. Due to the scoring being different for functional and symptom scales (higher scores are indicative of better health in the functional scales, while higher scores in the symptom scales are indicative of worse health), best and worst scores were considered differently depending on the type of scale. For the functional scales, those scoring 100 on the functional scales were considered to have the best scores, while those scoring 0 were considered to have the worst scores. It was the opposite for the symptom scales, meaning that those scoring 100 were considered to have the worst scores, while those scoring 0 were considered to have the best scores.

After the best and worst scores for each scale were tallied, frequency distribution tables and graphs were then constructed for each scale mentioned above. Because the symptom scales of the QLQ-LMC21 and QLQ-C30 are scored in the opposite manner of the functional scales of the SF-36 (a low score on a symptom scale is desirable and indicates good health and a low amount of symptoms, whereas a low score on the functional scale indicates poor health and low functioning), the scores needed to be transposed before a meaningful comparison could occur. The QLQ-LMC21 and QLQ-C30 scales scores were transposed and placed into one of ten "buckets". For example, a score of 0-10 for the QLQ-LMC21 or QLQ-C30 was assumed as being equivalent and comparable to a score of 90-100 on the SF-36. The ten buckets created were in intervals of 10: 0-10, 10-20, 30-40, 40-50, 50-60, 60-70, 70-80, 80-90, 90-100. Those scoring from 0-10 are those in the worst health, while those scoring 90-100 are those in the best health.

Once scores were transposed into their appropriate buckets, data was then plotted into Excel and graphs created to show any left or right skewness in the distribution of scale scores. Because there is no consensus on how to define floor and ceiling effects mathematically, it was determined *a priori* that floor and ceiling effects were present when scale scores were found between 0 and 10, and 90 and 100, respectively.

6 Results

The results will be structured in a similar way to the methods section. Results will be divided into three sections: (1) the translation process; (2) content validity, psychometric validity, and equivalence of the QLQ-LMC21; and (3) scale reliability, validity and responsiveness in comparison to the SF-36 and QLQ-C30.

6.1 Translation results

6.1.1 Patient characteristics

As part of the translation process, the intermediary version of the QLQ-LMC21 was pilot-tested on ten patients (eight males, two females) diagnosed with CRC liver metastases and being treated at Rikshospitalet in Oslo. All were native Norwegian speakers, with the exception of one native Swedish speaker. All patients completed the questionnaire and were interviewed in the hospital. They were between the ages of 57 and 77, with a mean age of 66 years. Characteristics are shown in Table 6.

Table 6. Translation patient characteristics

Patients (n)	10
Mean age (range)	66 (57-77)
Gender	
Men (%)	8 (80)
Women (%)	2 (20)

6.1.2 Forward translation

General comments regarding the translation

The words "have you" have been translated as "har du" in other questionnaires. If possible (grammatically), the same translation was chosen.

Items:

Item 33:

Original English version: ***Have you worried about losing weight?***

We felt that the FW2 version was more colloquial, straightforward, and written in a way that more people would understand.

Forward translation: ***Har du bekymret deg for å gå ned i vekt?***

Item 41:

Original English version: ***Have your skin or eyes been yellow (jaundiced)?***

The word "gulaktige" translates to "yellowish" which we agreed would be a more appropriate word in Norwegian than "gule", which translates to just "yellow."

Forward translation: ***Har huden eller øynene dine vært gulaktige (gulsott)?***

Item 44:

Original English version: ***Have you felt lacking in energy?***

We decided that the word "mangler" was more appropriate because it translates into "lacking", which is much less open to interpretation than the word "lite", which could be more widely interpreted by patients than "mangler."

Forward translation: ***Har du følt at du mangler energi?***

Item 46:

Original English version: ***Have you had trouble talking about your feelings to your family or friends?***

For consistency, the translation beginning with "har du" was chosen.

Forward translation: ***Har du hatt vanskeligheter med å snakke om dine følelser med familie eller venner?***

Item 47:

Original English version: ***Have you felt stressed?***

Both FW1 and FW2 had the same translation.

Forward translation: ***Har du følt deg stresset?***

Item 48:

Original English version: ***Have you felt less able to enjoy yourself?***

We felt the verb "å more" was a more suitable word in this context, as it is used in the reflexive to mean "to enjoy yourself." "Å nyte" is usually not used in the reflexive and used to mean general enjoyment. We decided "å more" would be more easily understood by patients in the context of the survey question.

Forward translation: ***Har du følt at du er mindre i stand til å more deg?***

Item 50:

Original English version: ***Were you worried about your family in the future?***

There was some question about the intended meaning of this question. After some discussion, we felt that this question was asking if the patient was worried about their family's future. Therefore, we chose the FW2 translation. FW2 literally translated into "Have you worried about your family's future", while FW1 translated into "Were you worried about your family in the future". FW1 is a direct translation of the English version, however we felt that FW2 was written in a clearer way that more patients would understand. We also chose to begin the question with "har du" to maintain consistency with the other items.

Forward translation: ***Har du vært bekymret for din families fremtid?***

Item 51:

Original English version: ***Has the disease or treatment affected your sex life (for the worse)?***

There was some discussion regarding the parenthetical phrase at the end of the translation. FW2 translated this to (til det verre), while FW1 incorporated the parenthetical phrase into the question (på en negativ måte). We decided that FW1 retained the original meaning of the English version, while sounding better in Norwegian.

Forward translation: ***Har sykdommen eller behandlingen påvirket sexlivet ditt på en negativ måte?***

6.1.3 Backward translation

In one case, after discussion between the backward translators, changes to the original forward translation were made because it was decided that a more appropriate word could be used to better capture the original meaning of the question.

Items:

Item 33:

Forward translation: ***Har du bekymret deg for å gå ned i vekt?***

Both backward translators translated this item back to the original English meaning, however, after some discussion, we decided that the word "vekttap" is the more appropriate word for unintentional weight-loss, rather than "å gå ned i vekt", which could be interpreted to mean someone on a diet who is unable to intentionally lose weight. We decided that the original intention of the question was to ascertain the level of worry about unintentional weight-loss, as that is a common affliction with cancer patients. The first intermediary Norwegian version was changed from "Har du bekymret deg for å gå ned i vekt?" to "Har du bekymret deg for

vektap?”

Change to forward translation: ***Har du bekymret deg for vekttap?***

Item 41:

Forward translation***: Har huden eller øynene dine vært gulaktige (gulsott)?***

BW1 chose “yellow-tinted” and BW2 chose “yellow” as a translation of "gulaktige". After a short discussion, we decided that “yellow-tinted” was a more direct translation of the Norwegian word “gulaktige” used in the Norwegian version of the item. It also makes more sense in Norwegian to describe skin as “yellow-tinted” rather than “yellow”. There was agreement that the backward translations were in agreement with the forward translations. No changes were made to the forward translation.

Item 44:

Forward translation: ***Har du følt at du mangler energi?***

BW1 used the phrase “felt at a loss of energy”, while BW2 used the phrase “experienced a lack of energy” as a translation of the forward translation. Through discussion, we decided that “felt at a loss of energy” better captured the original English item and better captures the perception of loss of energy. There was agreement that the backward translations were in agreement with the forward translations. No changes were made to the forward translation.

Item 46:

Forward translation: ***Har du hatt vanskeligheter med å snakke om dine følelser med familie eller venner?***

Both backward translations were similar to the original English text. There was agreement that the backward translations were in agreement with the forward translations. No changes were made to the forward translation.

Item 47:

Forward translation: ***Har du følt deg stresset?***

Both backward translations were similar to the original English text. There was agreement that the backward translations were in agreement with the forward translations. No changes were made to the forward translation.

Item 48:

Forward translation: ***Har du følt at du er mindre i stand til å more deg?***

Though both translations essentially captured the original English version, there were differing phrases used to express enjoyment. BW1 used the phrase “less able to have fun”, while BW2 used the phrase “less capable of enjoying yourself” as a translation for "mindre i stand til å more deg". “I stand”, which is used in the Norwegian item, can mean either “capable” or “able” in the Norwegian language. However, “å more” more directly translates into “to have fun”, so we decided that BW1 was slightly more true to the original English item. Also, we discussed the connotation of “ability” vs. “capability”, with ability implying being able or not able to complete a task for various reasons, while “capability” implies having the capacity to complete or not complete a task. There was agreement that the backward translations were in agreement with the forward translations. No changes were made to the forward translation.

Item 50:

Forward translation: ***Har du vært bekymret for din families fremtid?***

There was some discussion in the forward translation about the intended meaning of this question, and it was decided that it was “Have you been worried about your family’s future?” It was then translated into Norwegian according to this meaning. Both backward translations were identical to this text. There was agreement that the backward translations were in agreement with the forward translations. No changes were made to the forward translation.

Item 51:

Forward translation: ***Har sykdommen eller behandlingen påvirket sexlivet ditt på en negativ måte?***

There was some discussion in the forward translation about the parenthetical phrase at the end of this question. It was decided that the parenthetical phrase would be incorporated into the main sentence. It was then translated into Norwegian accordingly. Both backward translations were identical to this text. There was agreement that the backward translations were in agreement with the forward translations. No changes were made to the forward translation.

6.1.4 Feedback from EORTC

EORTC recommended 1 change to the backward translation (item 41), and 2 changes to the intermediary version of the questionnaire (items 50 and 51).

Backward translation recommended changes:

Item 41:

Original English Version: ***Have your skin and eyes been yellow (jaundiced)?***

Backward translation: ***Have your eyes or skin been yellow-tinted (jaundice)?***

Recommended change:

EORTC recommends that “skin” and “eyes” be switched in the sentence to match the original English version.

New backward translation item: ***Have your skin and eyes been yellow-tinted (jaundice)?***

Intermediary questionnaire recommended changes

Item 50:

Original English Version: ***Have you worried about your family in the future?***

Questionnaire item: ***Har du vært bekymret for din families fremtid?***

Recommended change:

EORTC advises that it would be better to say "family in the future" rather than "family's future" as the question asks about how the illness may indirectly affect patient`s family in the future rather than the future of the family - as in if family will break up or not.

New intermediary questionnaire item: ***Var du bekymret for din familie i fremtiden?***

Item 51:

Original English Version: ***Has the disease or treatment affected your sex life (for the worse)?***

Questionnaire item: ***Har sykdommen eller behandlingen påvirket sexlivet ditt på en negativ måte?***

Recommended change:

To remain consistent with the original English version, EORTC recommends putting “for the worse” in brackets, as in the English version.

New intermediary questionnaire item: ***Har sykdommen eller behandlingen påvirket ditt sexliv (til det verre)?***

These changes were implemented and the intermediary questionnaire was sent to EORTC for

proofreading by an outside translation agency.

6.1.5 Feedback from translation agency

The outside translation agency recommended three changes to the intermediary questionnaire, items 46, 49, and 51. No explanation for the changes was given and can only be inferred.

Item 46:

Questionnaire item: ***Har du hatt vanskeligheter med å snakke om dine følelser med familie eller venner?***

Recommended change: ***Har du hatt vanskeligheter med å snakke om følelsene dine med familie eller venner?***

This change is assumed to have been recommended because the translation agency felt that "følelsene dine" has a better flow in Norwegian than "dine følelser". Both have the same meaning but may sound better to a Norwegian speaker.

Item 49:

Questionnaire item: ***Har du vært engstelig for helsen din i fremtiden?***

Recommended change: ***Har du vært bekymret for din fremtidige helsetilstand?***

Question 49 was not part of forward and backward translation process because it was one of the pre-translated items furnished by EORTC. The translation agency used in this process decided to completely change this question from its previous form. They disagreed with the use of the word "engstelig" as a translation for the original English word "worried", the use of the word "helse" for a direct translation of the original English word "health", and decided to change the noun "fremtiden" into the adjective "fremtidige" as a translation for the phrase "future health". "Engstelig" means anxious, while "bekymret" is a direct translation of "worried". While both words can be interpreted as responses to some sort of threat, worry may be perceived as the thought process that leads to anxiety, so someone may be worried, but not necessarily anxious. Anxiety may also be perceived as a more severe state-of-mind than worry. Patients who answer this question when it contains the word "engstelig" may answer this question differently than they would if the item used the word "bekymret" because the words imply different emotional states. Though "helse" is a direct translation for the original English word "health", "helsetilstand" means health condition, which is implied by the use of the word "helse" in the original Norwegian item. The translation agency decided that it was better to be explicit rather than implied in the item.

Item 51:

Questionnaire item: ***Har sykdommen eller behandlingen påvirket ditt sexliv (til det verre)?***

Recommended change: ***Har sykdommen eller behandlingen påvirket sexlivet ditt (negativt)?***

Similar to the change recommended for item 46, the translation agency felt that "sexlivet ditt" had a better flow in Norwegian than "ditt sexliv". They also disagreed with the direct translation from the English item of the parenthetical phrase "til det verre", and simplified with the word "negativt".

These recommended changes were implemented in the first intermediary version of the QLQ-LMC21.

6.1.6 Pilot testing of the first intermediary version of QLQ-LMC21

Most patients reported no difficulties in completing the questionnaire. None of the questions were upsetting to the patients or contained difficult words. One patient expressed concern with questions 38 and 50. The word "kribling" in question 38 was noted as referring more to a tingling sensation than to the loss of sensation and numbness that the patient experiences as a side-effect of chemotherapy for the disease. This patient also noted for question 50 that he preferred the phrasing "Har du vært bekymret for din familie i fremtiden" over "Var du bekymret for din familie i fremtiden" because the former is consistent with the phrasing of the previous questions in the questionnaire. Another patient reported a difficulty with question 51. She was not certain if her age, disease, or treatment was affecting her sex-life and was uncertain how to answer the question.

6.1.7 Final acceptance of QLQ-LMC21 from EORTC

The translation report was updated with the results from the pilot testing and sent to EORTC for final review. EORTC requested the item 50 remain consistent with the English version and the item version "Var du bekymret…" was selected to remain in the final version of the instrument. EORTC noted the word "golsot" in question 41 as misspelled. The spelling was changed from "golsot" to "gulsott" and the final Norwegian translated version of the QLQ-LMC21 was accepted by the agency and made available in their database for widespread use by researchers.

6.1.8 Content validity

The multi-layered and iterative translation process identified and corrected a number of potential weaknesses of the instrument. For example, the backward translators were able to offer a better translation for item 33, which was ultimately accepted by EORTC and is now a

part of the final version of the questionnaire. Also, having four translators in agreement on most of the questions was also helpful in reinforcing the assumed quality of the questionnaire items. During pilot testing, most patients reported no difficulties in completing the questionnaire. None of the questions were upsetting to the patients or contained difficult words.

6.2 Content validity, psychometric validity and equivalence results

6.2.1 Patient characteristics

After pilot testing and final acceptance by the EORTC QoL Group, the QLQ-LMC21 was administered to 22 patients participating in the CoMet study. The mean age was 68, with patients ranging in age from 49-86. There were 12 men (54.5%) and 10 women (45.5) with an average time since primary CRC diagnosis of 10.4 months. Table 7 shows patient characteristics for the psychometric assessment.

Table 7. Psychometric assessment patient characteristics

Patients (n)	22
Mean age (range)	68 (49 - 86)
Gender	
Men (%)	12 (54.5)
Women (%)	10 (45.5)
Time since diagnosis (months) [a] (range)	10.4 (3 - 37)

[a] CRC diagnosis

6.2.1 Content validity

Content validity of the QLQ-LMC21 was explored using response distributions to see the non-response rate of each item. Response distributions are found in Table 17 in the appendix. Of the 21 items of the instrument, there were three items with missing responses. Items 35 and 37 each had one missing response (4.5%), while question 51, which asks about sexual function, had three missing responses (13.6%). Out of these three items, only item 51 was a newly translated item translated in this study. It was, however, not surprising that question 51 had a relatively high missing response rate because it asks about sexual habits. Patients may find questions about sex sensitive or offensive. In terms of patient acceptance, content validity is assumed to be good.

6.2.2 Psychometric validity

The psychometric validity of the QLQ-LMC21 was assessed in terms of item internal consistency, item convergent validity, and discriminant validity using PPMCC to correlate individual items to each scale. Table 8 displays the results in an MTMI matrix, while table 9 displays the summated results of item internal consistency and the convergent and

discriminant validity tests. Generally, the internal consistency for the items of all four scales was good, with all but two items exceeding the standard of .40 with significance. Item 42 from the abdominal pain scale had a PPMCC of -.01, and item 47 from the anxiety scale had a PPMCC of .38. However, neither of the scores was significant.

The instrument performed well in terms of item discriminant validity. In all but four correlations (item 42 and item 44), the item in question scored significantly higher (>2SE, where the SE of the correlation matrix was .06, meaning two standard errors above was .12.) with its own scale than with other scales. These results may indicate that there may be some construct overlap between item 42 and the three other scales, and item 44 and the abdominal pain and eating problems scales.

Table 9 shows the results for the item convergent and discriminant validity tests. The activity/vigor and eating problems scales scored 100% in tests of item convergent validity, meaning all items in the scale scored above a .40. In the abdominal pain scale, two of the three items scored above .40, giving a convergent validity success rate of 66.7%. In the anxiety scale, 3 out of 4 items scored above .40, giving a success rate of 75%.

The discriminant validity tests were just slightly more successful than the convergent validity tests. However, as with the convergent validity tests, only two of the four scales scored 100% success. The eating problems and anxiety scales scored 100%, meaning all item-correlations in the scales were significantly higher (>2SE) than the correlations between the items and the other scales. Due to item 42, the abdominal pain scale had only 6 of 9 significantly higher correlations, giving it the lowest rate at 66.7%. The activity/vigor scale had 7 of 9 significantly higher correlations, giving it a success rate of 77.8%. Item 44 was problematic and scored within 2 SE of its own item-scale correlation with the abdominal pain and eating problems scales.

It is important to note that due to a small sample size, the SE will be higher, thus lessening the likelihood of supporting item discriminate validity (Fayers, 2005).

Table 8. QLQ-LMC21 Item means with standard deviations and their Pearson correlations with scales

					QLQ-LMC21 Symptom Scales			
					Abdominal		Eating	
Scale	Item	Mean	SD	N	Pain	Activity/vigor	Problems	Anxiety
Abdominal Pain	39	1,47	0,84	19	**,58****	,40	,34	-,02
	40	1,56	0,84	19	**,75****	,49*	,31	,16
	42	1,32	0,75	19	**-,01**	,48*	,02	,29
Activity/Vigor	37	2,28	0,83	18	,49*	**,67****	,43	0,53*
	43	2,16	0,83	19	,64*	**,75****	,50*	,50*
	44[a]	2,32	0,75	19	,51*	**,55****	,46*	,04
Eating Problems	31	1,21	0,55	19	,30	,48*	**,75****	-,07
	32	1,42	0,69	19	,26	,55*	**,75****	-,16
Anxiety	47[a]	1,16	0,37	19	-,03	-,15	-,26	**,28**
	48[a]	1,74	0,87	19	,42	,50*	-,19	**,55****
	49	2,11	0,74	19	,12	,34	,13	**,70****
	50[a]	1,79	0,86	19	,04	,28	-,15	**,64****

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Underlined values corrected for overlap

Standard error of the correlation matrix = .06

[a] translated item

Table 9. Psychometric properties of the QLQ-LMC21

Scale	# items	Item internal consistency [a]	Item discriminant validity [b]	Item convergent validity test [c]	Item discriminant validity test [d]
Abdominal Pain	3	(-).01 - .75	(-).02 - .49	66,7	66,7
Activity/vigor [e]	3	.55 - .75	.04 -.64	100	77,8
Eating Problems	2	,75	(-).07 - .48	100	100
Anxiety [e]	4	.28 - .70	(-).26 - .42	75	100

[a] Range of correlations between items and hypothesized scale corrected for overlap

[b] Range of correlations between items and other scales

[c] Item convergent validity scaling success (%) i.e. number of item-scale correlations greater than ,40/total number of correlations (corrected for overlap)

[d] Item discriminant validity scaling success (%) i.e. number of correlations of items with own scales significantly higher (> 2SE) than correlations with other scales/total number of correlations

[e] constains newly translated items

6.2.3 Equivalence

The standard for equivalence of the Norwegian language QLQ-LMC21 with the English language version was considered to be satisfactory when there was a high degree of patient acceptance of the measure, a within item-scale PPMCC of .40 for items corrected for overlap, and between-scale scores lower than two SE below the within item-scale correlation was

achieved. The instrument had a high degree of patient acceptance, with two items missing one response each, and one item missing three responses. Table 9 displays these results of the item convergent and discriminant validity tests. Two scales had convergent validity scaling failures, the abdominal pain scale and the anxiety scale, with item 42 at a -.01 and item 47 at .28. Neither score, however, has statistical significance. Two scales also had divergent validity scaling failures, the abdominal pain scale and the activity/vigor scale. Item 42 again caused the scaling failures for the abdominal pain scale, and item 44 of the pain scale scored within two SE of the correlation for its own scale.

6.3 Reliability, validity and responsiveness results

6.3.1 Internal consistency reliability

Reliability of the QLQ-LMC21 scales was assessed using Cronbach's alpha. Results can be found in Table 10. Both between-scale alphas and overall scale alphas were calculated. The recommended standard of .70 was used to evaluate the reliability of the scales, with between scale scores below .70 desirable, but overall scale scores above .70 desirable. None of the between-scale scores were above .70. There were three of the four scales that scored an overall alpha of above .70, with the abdominal pain scale scoring .57.

Table 10. Correlations between QLQ-LMC21 scales and internal consistency using Cronbach's alphas

	Correlation between scales				
Scales	Abdominal Pain	Activity/vigor	Eating Problems	Anxiety	Cronbach's alphas
Abdominal Pain	1	,62	,30	,20	,57
Activity/vigor [a]	,62	1	,55	,39	,84
Eating Problems	,30	,55	1	-,12	,85
Anxiety[a]	,20	,39	-,12	1	,74

[a] contains newly translated items

6.3.2 Reliability estimates of comparable scales

Table 11 shows the exploration of the reliability of the corresponding scales of the SF-36, QLQ-C30, and QLQ-LMC21. The scales and questions compared can be found in Table 16 in the appendix. The reliabilities of the pain scales, vitality/fatigue/vigor scales, and mental health/emotional role-functioning/anxiety scales were calculated separately and in the case of the QLQ-LMC21 and QLQ-30, combined to explore if the reliability of the scales was increased or decreased when combined.

The means for the scales of the QLQ-C30 and QLQ-LMC21 scales were similar. The biggest difference in means was between the emotional role functioning scale of the QLQ-C30 and

the anxiety scale of the QLQ-LMC21 at 1.3 and 1.6, respectively, for a difference of .3. The range for the QLQ instruments is three, meaning the patient only has three response choices and makes for more compact means. The bodily pain scale of the SF-36 has a range of five, while the vitality and mental health scales have a range of four. Additionally, higher means for the SF-36 scales indicate better health. Lower scores on the QLQ instruments indicate better health, meaning that scores closer to 1 indicate better health.

All scales scored above a .70 alpha with the exception of the abdominal pain scale of the QLQ-LMC21, which scored a .57.

Table 11. Reliability estimates of comparable scales of the SF-36, QLQ-LMC21, and QLQ-C30

Content area and source	No. of items	Mean	St. Dev	Cronbach's alpha
Bodily pain/pain/abdominal pain				
SF-36	2	2,2 [a]	2.5	.85
QLQ-C30	2	1.5	1.7	.83
QLQ-LMC21	3	1.5	1.9	.57
QLQ-LMC21 + QLQ-C30	5	1.5	3.3	.79
Vitality/fatigue/vigor				
SF-36	4	3,2 [b]	3.5	.81
QLQ-C30	3	2.0	1.8	.84
QLQ-LMC21	3	2.3	2.4	.84
QLQ-LMC21 + QLQ-C30	6	2.1	3.4	.88
Mental health/emotional role functioning/anxiety				
SF-36	3	4,4 [b]	3.2	.85
QLQ-C30	4	1.3	2.4	.93
QLQ-LMC21	4	1.6	2.3	.74
QLQ-LMC21 + QLQ-C30	8	1.5	4.5	.90

Higher SF-36 means indicate better health
Higher QLQ-C30 and QLQ-LMC21 means indicate worse health
All QLQ mean scores based on a Likert scale with 4 answer choices
a Based on a Likert scale with 6 answer choices
b Based on a Likert scale with 5 answer choices

6.3.2 Concurrent validity

Concurrent validity was assessed using a MTMM matrix. Results can be found in Table 12. The same three corresponding scales of the SF-36, QLQ-C30, and QLQ-LMC21 that were examined in the reliability analysis were compared here: the pain scales, vitality/fatigue/vigor scales, and mental health/emotional role-functioning/anxiety scales. All instruments scored above .40 with statistical significance in the (convergent) validity diagonals. Discriminant

validity was also unanimous, with no correlations in the triangles being higher than in the validity diagonals.

For the QLQ-LMC21 and the SF-36, all correlations in the validity diagonal were above .70 and statistically significant, giving evidence for convergent validity. Correlations in the triangles were lower than the correlations in the validity diagonals, giving evidence for discriminant validity between these instruments for these three scales. However, three of the six correlations in the triangles were not found to be statistically significant.

For the QLQ-LMC21 and the QLQ-C30, the correlations in the validity diagonals were above .6 and significantly significant, giving evidence for convergent validity. All correlations in the triangles were lower than those in the validity diagonals, also providing evidence for discriminant validity. As with the scales for the QLQ-LMC21 and SF-36, three of the six correlations in the triangles were not found to be statistically significant.

The QLQ-C30 and SF-36 scored both the highest correlation in any of the validity diagonals (.91) and the highest correlation in any of the triangles (.68). The emotional role functioning scale of the SF-36 and the mental health scale of the SF-36 had a statistically significant correlation of .91, while the pain scale of the QLQ-LMC21 and the mental health scale of the SF-36 had a statistically significant correlation of .68, Also, three of the six correlations in the triangles were not significantly significant. These results, however, provide evidence for both convergent and discriminant validity.

6.3.3 Responsiveness

Responsiveness in the form of floor and ceiling effects was explored by first analyzing the best and worst possible scores for the three comparable scales of the QLQ-LMC21, QLQ-C30, and SF-36. Results are found in Table 13. Floor effects (percent with minimum score) ranged from 0%-4.5%. The pain scale of the QLQ-C30 was the only measure where a patient received the lowest score possible. Ceiling effects (percent with maximum score) were much more varied and ranged from 9.1%-59.1%. The emotional role scale of the QLQ-30 had the highest percentage of best possible scores at 59.1%. The activity/vigor scale of the QLQ-LMC21, fatigue scale of the QLQ-C30, and vitality scale of the SF-36 each had the lowest percentage at 9.1%.

Table 12. Concurrent validity in a MTMM matrix using Pearson correlation coefficients between scales from the SF-36, QLQ-C30 and QLQ-LMC21

		SF-36			QLQ-C30			QLQ-LMC21		
		Bodily pain	Vitality	Mental health	Pain	Fatigue	Role Functioning - Emotional	Abdominal pain	Activity/vigor	Axiety
SF-36	Boldily pain	1	,57**	,68**						
	Vitality	,57**	1	.28						
	Mental health	,68**	.28	1						
QLQ-C30	Pain	**,86****	**,47***	**,68****	1	.31	,56**			
	Fatigue	**.33**	**,78****	**.17**	.31	1	.03			
	Role Functioning - Emotional	**,47***	**.10**	**,91****	,56**	.03	1			
QLQ-LMC21	Abdominal pain	**,72****	**,53***	**.27**	**,66****	**,52***	**0.14**	1	,62**	.20
	Activity/vigor	**,52***	**,80****	**.33**	**,51***	**,85****	**.22**	,62**	1	.39
	Anxiety	**,55****	**.25**	**,77****	**,56****	**.18**	**,79****	.20	.39	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 13. Floor and ceiling effects - best and worst possible score percentages of comparable QLQ-LMC21, QLQ-C30, and SF-36 scales (N=22)

Scales	Mean (SE)	Worst possible score % (n)	Best possible score % (n)
QLQ-LMC21 Symptom Scales (0-100, 100 worst health)			
Abdominal Pain	18.1 (4.5)	0	45.5 (10)
Activity/vigor	42.9 (5.7)	0	9.1 (2)
Anxiety	22.3 (4.1)	0	13.6 (3)
QLQ-C30 Symptom Scales (0-100, 100 worst health)			
Pain	17.4 (6.0)	4.5 (1)	54.5 (12)
Fatigue	36.1 (4.7)	0	9.1 (2)
QLQ-C30 Functional Scale (0-100, 0 worst health)			
Role Functioning - Emotional	89.8 (4.3)	0	59.1 (13)
SF-36 Functional Scales (0-100, 0 worst health)			
Bodily Pain	73.8 (6.0)	0	40.9 (9)
Vitality	57.0 (4.6)	0	9.1 (2)
Mental Health	87.2 (3.3)	0	22.7 (5)

Floor and ceiling effects for the comparable scales are also shown in Figures 5-7. Scores have been rescaled from 0-100, with 0 representing the worst health. There appear to be ceiling effects in the pain (Figure 5) and mental health (Figure 7) scales of the measures, with the vast majority of patients scoring in the top quartile of scores of all measures. The vitality/fatigue (Figure 6) scales appear to be a bit more normally distributed, with a majority of respondents scoring between 50 and 70. The anxiety/emotional role-functioning/mental health distributions (Figure 7) showed some tendencies for ceiling effects in the QLQ-C30 and SF-36. The QLQ-C30 had 16 respondents and the SF-36 12 respondents achieve a score between 90 and 100, while the QLQ-LMC21 only had 3 respondents achieve the same. Half of patients (11) scored 80-90 points with the QLQ-LMC21, but on the whole patient scores were much more distributed than with the other mental health scales, with eight patients ranging in score from 10 - 80. The SF-36 only had two patients score below an 80, and the QLQ-C30 had three patients with a score below 80.

Figure 5. Frequency distributions of QLQ-LMC21, QLQ-C30, and SF-36 pain scales

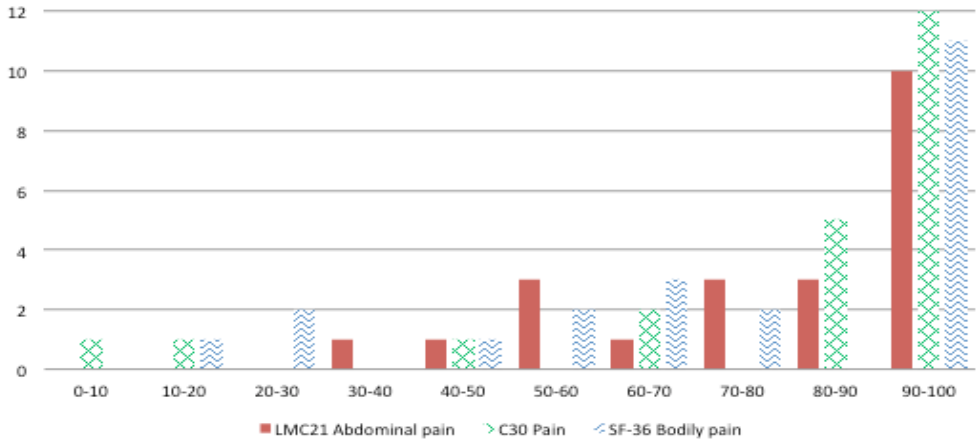


Figure 6. Frequency distributions of QLQ-LMC21, QLQ-C30, and SF-36 vitality/fatigue scales

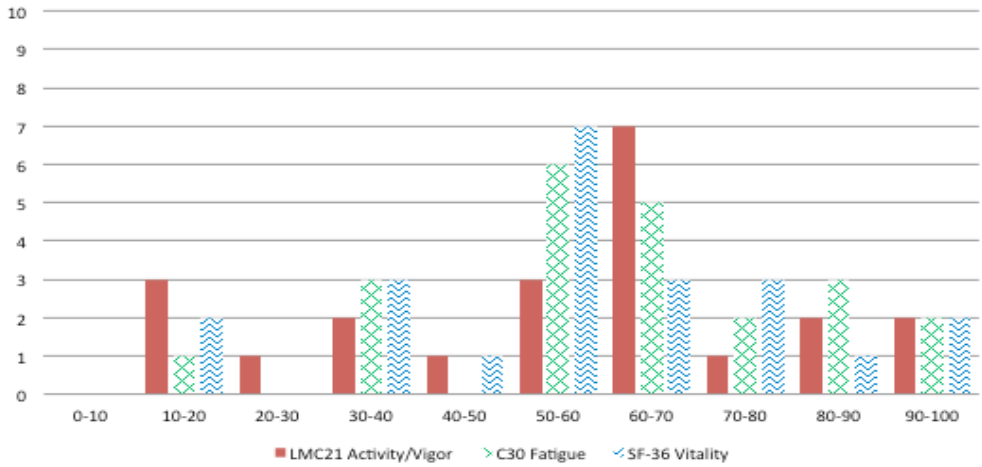
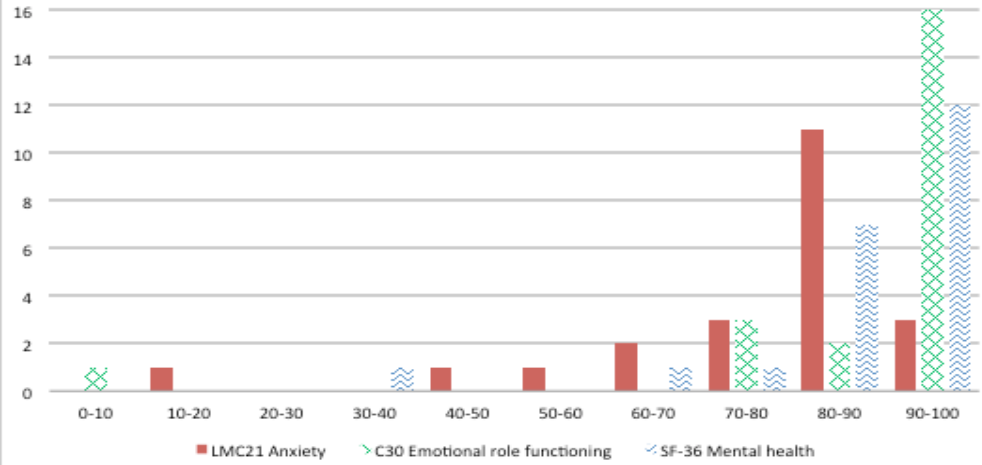


Figure 7. Frequency distributions of QLQ-LMC21, QLQ-C30, and SF-36 mental health scales



7 Discussion

7.1 Study objectives

This study aimed to explore the methods used in the translation and psychometric assessment of HRQoL instruments and preliminarily assess the quality of the Norwegian QLQ-LMC21 in terms of reliability, validity, responsiveness, and equivalence with the English version. Both the translation and psychometric assessment were approached in a methodical, systematic and transparent way to increase the likelihood of developing a high quality instrument and subsequently accurately assessing its quality.

7.2 Main findings

7.2.1 Translation process

The iterative nature of the forward and backward translation process was useful in finding weaknesses in the wording of questions that could potentially affect the content validity of the instrument and helped to improve the equivalence between the English and Norwegian versions. Content validity of the questionnaire as a result of the translation process, including pilot testing, is assumed to be good.

The EORTC QoL Group was unwilling to accept any items on the Norwegian QLQ-LMC21 that deviated in structure from the original English version, even if, arguably, the Norwegian item was made to be more clear and understandable to patients. For example, item 51 was modified by removing a parenthetical phrase and seamlessly incorporating it into the sentence. It was felt by all translators that when left with the parenthetical phrase, the sentence was redundant, and removing the parenthetical phrase was an effective way to improve the flow and clarity of the question without compromising conceptual, functional, linguistic equivalence with the English version of the questionnaire. EORTC, however, wishes to maintain absolute consistency with the English version to ensure equivalence. They may also be hesitant to accept structural changes to translated versions to avoid making changes to the original English versions.

Another occurrence of note was the recommended change of item 49 by the outside translation agency during the translation process, as item 49 had previously been translated for another Norwegian language EORTC instrument and was provided to this study already translated. This highlights the subjective nature of the translation process and how it is very easy and possible to have translated instruments that do not maintain equivalence or consistency with the original version.

7.2.2 Psychometric assessment

Item 51 (the sexual function item) was the only translated item to have problems with missing responses. As it is not unusual for patients to have difficulties with questionnaire items regarding sex, it is assumed that the translated measure has good patient acceptance and good content validity. In a study of cancer patient HRQoL by Fairclough and Cella in the US, the response rate for the item "I am satisfied with my sex life" was only 7% (Fairclough, 1996). The same study had a general non-response rate that ranges for each question from 0% to 12%. Additionally, patients may be unsure whether to attribute any difficulties in their sex lives to the disease, treatment, other comorbid illnesses, or a consequence of being older in age and may simply choose to not answer the question.

Generally, the psychometric quality of the QLQ-LMC21 was fair to good. Internal consistency was fair to good, with 10 out of 12 item correlations scoring above .40 within their own scales. Item discriminant validity tests fared just slightly better than item convergent validity tests, meaning that while the scales may not be measuring what they purport to be measuring, they may be more likely to be measuring discrete concepts. Item 42 in the abdominal pain scale performed rather poorly and caused scaling failures in both the convergent and discriminant scaling tests. Item 42 asks if the patient has had back pain, which may be a peripherally associated symptom of liver metastases. However, it is possible that these patients attributed peripheral pain to their stomach region rather than their back and do not perceive the pain as back pain, causing this item to correlate poorly with the abdominal pain scale.

Equivalence with the English version was acceptable. The instrument had high patient acceptance and fair to good results in the item convergent and discriminant validity tests, with item 42 causing a majority of the failures in this area.

Scale reliability was fair, with one of the four scales (the abdominal pain scale) having a Cronbach's alpha below .70. The low reliability of the abdominal pain scale may be a consequence of item 42's poor performance. The two scales that contained newly translated items (activity/vigor and anxiety), however, scored .84 and .74, good and acceptable, respectively.

The combined reliability estimates for the QLQ-LMC21 and QLQ-C30 were good, with all combined scores being over .70. The abdominal pain scale of the QLQ-LMC21 was improved with the addition of the items from the QLQ-C30 pain scale score, helping to

overcome the negative effects of item 42. However, the combined reliability of the scales was less than the alpha (.83) of the QLQ-C30 pain scale alone, indicating that the QLQ-C30 may be more reliable alone than in conjunction with the QLQ-LMC21 abdominal pain scale. The reliabilities of the fatigue/vigor scales of the QLQ-C30 and QLQ-LMC21 were both improved to .88 when combined, indicating that these scales work better together. The emotional role functioning scale of the QLQ-C30 was very high at .93, while the anxiety scale of the QLQ-LMC21 was .74. When combined with the anxiety scale of the QLQ-LMC21, the alpha dropped to a score of .90. Usually a drop in alpha with the addition of items is undesirable. Additionally, the reliabilities for the SF-36 were above .80 for all three scales, potentially pointing to the quality of the SF-36 and confirming the assumptions of this study that the SF-36 can be used as a stand-in for a "gold standard" instrument.

The concurrent validity of the three measures was good, with all measures showing a high degree of both convergent and divergent validity with each other. The QLQ-LMC21 performed well against all measures in terms of convergent validity, providing evidence that the instrument is measuring the constructs it purports. However, the number of statistically insignificant correlations in the triangles makes it difficult to draw conclusions about the discriminant validity of the instrument.

Ceiling effects appear to be present for two of the three scales compared, the pain and mental health scales, with the fatigue/vitality scores being more normally distributed. The most surprising finding was in the anxiety scale of the QLQ-LMC21. Only three patients scored the highest possible on this scale, as opposed to 16 for the QLQ-C30 emotional role-functioning scale and 12 for the SF-36 mental health scale. When thinking about scales that are assumed to measure similar constructs, one would assume that the distribution of responses would be similar between instruments, meaning that if ten people scored the highest score possible on one scale, the same ten people would in theory score the highest possible score on the comparable scale. That this did not happen for these particular scales could indicate that the QLQ-LMC21 is more able to discriminate between patients in the better health states for this scale, which would be congruent with the assumption that disease-specific measures are better at measuring small, yet significant changes in patient health.

Despite the methodological approach taken with the psychometric assessment, the sample size and time horizon make it difficult to draw many conclusions about the quality of the

Norwegian translated QLQ-LMC21 because it is assumed that some of these results may be the result of a small sample size and would differ significantly with a larger sample size. If, however, these findings were to be duplicated in a larger study, I would be hesitant to recommend the use of the instrument due to the poor performance of the abdominal pain scale. Nonetheless, results of convergent and concurrent validity, reliability of all scales other than the abdominal pain scale, and responsiveness are positive and very promising.

7.3 Limitations

As with any research, there are limitations with this study and conclusions must be drawn carefully. The small sample size (22 patients) makes it difficult to make definitive conclusions because of the possibility that this data is non-normally distributed and not representative of this patient group as a whole. Additionally, small sample sizes can affect Cronbach's alpha calculations, leading to incorrect reliability data.

Collecting HRQoL measurements from patients over several time points and statistically analyzing the results was outside the scope of this study due to the time horizon (4 months). Because of this, responsiveness was explored visually using floor and ceiling effects. It may, however, be difficult to draw definitive and statistically sound conclusions regarding an instrument's ability to measure change using this method.

The ability to address the quality of the single symptom items (because psychometric assessment is only concerned with scaling assumptions and leaves single items ignored) was also outside the scope of this study due to the time horizon and study design. The performance of single symptom items is, however, important, and was largely unaddressed in this study with the exception of the content validity exploration.

Additionally, the QLQ-LMC21 has not yet been studied extensively in its original English form, making comparisons and judgments about its quality even more difficult.

7.4 Further studies/research

Further research with a longitudinal study design and larger sample size could address the limitations present in this study, as well as several specific issues that were raised during the analysis. More research will allow more meaningful judgments to be made regarding the quality of the Norwegian version, the original English version, and other translated versions that may be developed in the future.

Additional tests of the abdominal pain scale with a larger sample size could help to clarify problems with item 42. Though item 42 performed very poorly in this analysis, it is not possible to make a definitive statement regarding its quality and appropriateness in the abdominal pain scale. A more robust study, however, may be able to explore these issues more fully, and possibly gather evidence for the removal of the item from the scale.

The quality of single symptom items was outside the scope of this study, but it could be explored in a longitudinal study where both between-groups and within-groups differences could be measured. Their reliability could be assessed using the test-retest correlation, as was done by Wan et al. in their validation of the Chinese QLQ-C30 (Wan, 2008). Their quality could be further explored by using a between-groups (for example surgery groups) test to assess the ability of the items to accurately differentiate between groups and their ability to measure change over time (responsiveness), as was done by Tan et al. in their validation of the QLQ-C30 and breast cancer specific EORTC module in Singapore (Tan, 2014) and Wan et al. in 2008 (Wan, 2008), respectively.

A longitudinal study using between groups and within group comparisons would also allow for more robust tests of responsiveness, such as effect size, standardized response mean, and the responsiveness statistic. The responsiveness of disease-specific instruments may be the biggest advantage they have over other instruments, therefore additional assessment of the responsiveness of the QLQ-LMC21 is important to determine whether or not it is better able to measure small changes in the HRQoL of CRC patients with liver metastases than other instruments, such as the QLQ-C30 or SF-36.

The results of this study suggest that it may be possible to supplement the SF-36 with the QLQ-LMC21 for patients with CRC liver metastases, without using the core EORTC cancer measure, the QLQ-C30. The SF-36 serves an important function because it can be used to derive QALYs in an economic evaluation. The QLQ-LMC21 is also important because it may offer a greater degree of responsiveness to the health states of this patient group. The function of the QLQ-C30 is, however, less clear. Some may argue that it is unnecessary or even unethical to burden patients already suffering from pain, fatigue, and loss of function to fill out multiple questionnaires that may be redundant or even yield invalid or unreliable results. Furthermore, asking patients to fill out multiple questionnaires may lead to patients becoming fatigued or overloaded and providing inaccurate or incomplete data, leading to poor and counterproductive data quality. Further studies of the psychometric qualities of the QLQ-LMC21 will cast more light on this subject.

8 Conclusion

This study has offered a transparent glimpse into the rigorous process of translating and checking the quality of the QLQ-LMC21 for use in patients with CRC liver metastases in Norway. Additionally, outside of the initial validation study performed by Blazeby et al. in 2009 (Blazeby et al., 2009), this is currently the only study that has explored and assessed the psychometric properties of the QLQ-LMC21. This study can serve as a starting point for further evaluation of the quality of the Norwegian QLQ-LMC21 and its ability to measure HRQoL in patients with CRC liver metastases, as well as to generally add to the limited worldwide body of knowledge for this instrument. With CRC rates rising for men and women, this instrument can serve an important role in evaluating interventions for this disease that improve HRQoL and extend patient lives.

References

Aaronson, N., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez., N., Filiberti, A., Osoba, D., Sullivan, M., . (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute, 85*(5).

Aaronson, N., Cull, A., Kaasa, S., Sprangers, M. (1994). A modular approach to quality of life assessment in oncology. *International Journal of Mental Health, 23*(2).

Abdalla, E. K., Vauthey, J.-N., Ellis, L. M., Ellis, V., Pollock, R., Broglio, K. R., . . . Curley, S. A. (2004). Recurrence and Outcomes Following Hepatic Resection, Radiofrequency Ablation, and Combined Resection/Ablation for Colorectal Liver Metastases. *Ann Surg, 239*(6), 818-827. doi: 10.1097/01.sla.0000128305.90650.71

Apolone, G., Filiberti, S., Ruggiata, R., Mosconi, P. (1998). Evaluation of the EORTC QLQ-C30 questionnaire: A comparison with SF-36 health survey in a cohort of Italian long-survival cancer patients. *Annals of Oncology, 9*, 549-557.

Augestad, L. A., Rand-Hendriksen, K., Stavem, K., & Kristiansen, I. S. (2013). Time trade-off and attitudes toward euthanasia: implications of using 'death' as an anchor in health state valuation. *Qual Life Res, 22*(4), 705-714. doi: 10.1007/s11136-012-0192-9

Beaton, D., Bombarier, C., Guillemin., F., Bosi Ferraz, M. (2000). Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures. *SPINE, 25*(24), 3186-3191.

Bensing, J. (2000). Bridging the gap. The separate worlds of evidence-based medicine and patient-centered medicine. *Patient Education and Counseling, 39*, 17-25.

Bergman, B., Aaronson, N.K., Ahmedzai, S., Kaasa, S., Sullivan, M. (1994). The EORTC QLQ-LC13: a modular supplement to the EORTC core quality of life questionnaire (QLQ-C30) for use in lung cancer clinical trials. *European Journal of Cancer, 30A*(5), 635-642.

Blazeby, J. M., Avery, K., Sprangers, M., Pikhart, H., Fayers, P., & Donovan, J. (2006). Health-related quality of life measurement in randomized clinical trials in surgical oncology. *J Clin Oncol, 24*(19), 3178-3186. doi: 10.1200/JCO.2005.05.2951

Blazeby, J. M., Fayers, P., Conroy, T., Sezer, O., Ramage, J., Rees, M., & European Organization for Research Treatment of Cancer Quality of Life, G. (2009). Validation of the European Organization for Research and Treatment of Cancer QLQ-LMC21 questionnaire for assessment of patient-reported outcomes during treatment of colorectal liver metastases. *Br J Surg, 96*(3), 291-298. doi: 10.1002/bjs.6471

Borgan, J. K. (2013). Dødsårsak, 2012 (Vol. 2015): Statistisk Sentralbyrå

Brazier, J. E., Harper, R., Munro, J., Walters, S.J., Snaith, M.L. (1999). Generic and condition-specific outcome measures for people with osteoarthritis of the knee. *Rheumatalogy*(38), 870-877.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*(3), 185-216.

Campbell, D., Fiske, D. (1959). Convergent and discriminant validation by multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Claasen, J. (2005). The gold standard: not a golden standard. *BMJ, 330*(7500), 1120-1121. doi: 10.1136/bmj.38449.476759.AE

Crocker L., A., J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich.

DeVon, H., Block, M., Moyle-Wright, P., Ernst, D., Hayden, S., Lazzara, D., Savoy, S., Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship, 39*(2), 155-164.

Dewolf, L., Koller, M., Velikova, G., Johnson, C., Scott, N., Bottomley, A.,. (2009). EORTC Quality of Life Group Translation Procedure.

Drummond, M. F., Sculpher, M.J., Torrance, G.W., O'Brien, B.J., Stoddart, G.L. (2005). *Methods for the Economic Evaluation of Health Care Programmes* (Third ed.). Great Britain: Oxford University Press.

EORTC. (2001). *EORTC Scoring Manual*. Brussels, Belgium.

EORTC. (2015). Retrieved 15/1-2015, 2015, from http://groups.eortc.be/qol/eortc-qlq-c30

Fairclough, D. L., Cella, D.F. (1996). Functional assessment of cancer therapy (FACT-G): non-response to individual questions. *Qual Life Res, 5*, 321-329.

Fayers, P., Hays, R. (2005). *Assessing Quality of Life in Clinical Trials* (2nd ed.): Oxford University Press.

Feeny, D., Eckstrom, E., Whitlock, P., Perdue, L. (2013). A Primer for Systematic Reviewers on the Measurement of Functional Status and Health-Related Quality of Life in Older Adults. *Agency for Healthcare Research and Quality (US)*.

Ferlay, J., Shin, H.R., Bray, F., et al. (2010). Cancer Incidence and Mortalitity Worldwide. *IARC CancerBase No.10.* Retrieved 10.5.2015, 2015, from http://globocan.iarc.fr

Ferrans, C. E., Zerwic, J. J., Wilbur, J. E., & Larson, J. L. (2005). Conceptual Models of Health-Related Quality of LIfe. *Journal of Nursing Scholarship*, 336-342.

Fretland, A. A., Kazaryan, A. M., Bjornbeth, B. A., Flatmark, K., Andersen, M. H., Tonnessen, T. I., . . . Edwin, B. (2015). Open versus laparoscopic liver resection for colorectal liver metastases (the Oslo-CoMet study): study protocol for a randomized controlled trial. *Trials, 16*(1), 73. doi: 10.1186/s13063-015-0577-5

George, D., Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference* (4th ed.). Boston: Allyn & Bacon.

Glick, H. A., Doshi, J.A., Sonnad, S.S., Polsky, D. (2015). *Economic Evaluation in Clinical Trials* (Second ed.). New York, New York: Oxford University Press.

Guillemin, F., Bombardier, C., Beaton, D. (1993). Cross-cultural Adaptation of Health-Related Quality of Life Measures: Literature Review and Proposed Guidelines. *Clinical Epidemiology, 46*(12), 1412-1432.

Guyatt, G., Feeny, D., Patrick, D. (1993). Measuring Health-realted Quality of Life. *Basic Science Review, 118*, 622-629.

Guyatt, G., Jaeschke, R., Singer, J. (1989). Measurement of health status: Ascertaining the Minimal Clinically Important Difference. *Controlled Clinical Trials, 10*, 407-415.

Guyatt, G., Walter, S., Norman, G. (1985). Measuring change over time: Assessing the usefuless of evaluative instruments. *Journal of Chronic Disease, 40*(2), 171-178.

Hambleton, R. (1993). Translating achievment tests for use in cross-national studies. *International Association for the Evaluation of Educational Achievment*.

Hays, R. D., Anderson, R., Revicki, D. (1993). Psychometric considerations in evaluating health-related quality of life measures. *Qual Life Res, 2*, 441-449.

Herdman, M., Fox-Rushby, J., Badia, X. (1998). A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. *Quality of Life Research, 7*, 323-335.

Hobart, J. C., Williams, L. S., Moran, K., & Thompson, A. J. (2002). Quality of Life Measurement After Stroke: Uses and Abuses of the SF-36. *Stroke, 33*(5), 1348-1356. doi: 10.1161/01.str.0000015030.59594.b3

Hviding, K., Juvet, L. K., Vines, D., & Fretheim, A. (2008). Colorectal cancer screening – effect on mortality and incidence rate of colorectal cancer.: Norwegian Knowledge Centre for Health Services.

ISOQOL. (2015). Goal and Purpose of HRQOL Measurement. *What Is Health-Related Quality of Life Research?* Retrieved 2.3.2015, 2015, from http://www.isoqol.org/about-isoqol/what-is-health-related-quality-of-life-research

Kavadas, V., Blazeby, J. M., Conroy, T., Sezer, O., Holzner, B., Koller, M., & Buckels, J. (2003). Development of an EORTC disease-specific quality of life questionnaire for use in patients with liver metastases from colorectal cancer. *European Journal of Cancer, 39*(9), 1259-1263. doi: 10.1016/s0959-8049(03)00236-3

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *Am J Health Syst Pharm, 65*(23), 2276-2284. doi: 10.2146/ajhp070364

Kuenstner, S., Langelotz, C., Budach, V., Possinger, K., Krause, B., Sezer, O. (2002). The comparability of quality of life scores: a multitrait multimethod analysis of the EORTC QLQ-C30, SF-36 and FLIC questionnaires. *European Journal of Cancer, 38*, 339-348.

Lenert, L., Kaplan, R. (2000). Validity and Interpretation of Preference-Based Measures of Health-Related Quality of Life. *Medical Care, 38*(9), 138-150.

Loge, J., H., Kaasa, S., Hjermstad, M.J., Kvein, T. (1998). Translation and performance of the Norwegian SF-36 health survey in patients with rheumatoid arthritis. Data quality, scaling assumptions, reliability, and construct validity. *J Clin Epidemiol, 51*(11), 1069-1076.

Magaji, B. A., Moy, F. M., Roslani, A. C., Sagap, I., Zakaria, J., Blazeby, J. M., & Law, C. W. (2012). Health-related quality of life among colorectal cancer patients in Malaysia: a study protocol. *BMC Cancer*.

Mutebi, A., Brazier, J., Walters, J. (2011). A comparison of the discriminative and evaluative properties of the SF-36 and the SF-6D index. *Qual Life Res, 20*(9), 1477 - 1486. doi: 10.1007/1s1136-101-988-1z

Nunnally, J. (1978). *Psychometric Theory* (2nd ed.). New York, New York: McGraw-Hill

Osoba, D. (2011). Health-related quality of life and cancer clinical trials. *Therapeutic Advances in Medical Oncology, 3*(2), 57-71. doi: 10.1177/

1758834010395342

Øynes, J., Brathaug, A.L. (2015). Health accounts 2014. Retrieved 8.5.2015, 2015, from http://www.ssb.no/en/nasjonalregnskap-og-konjunkturer/statistikker/helsesat

Patrick, D., Deyo, R. (1989). Generic and Disease-Specific Measures in Assessing Health Status and Quality of Life. *Medical Care, 27*(3), 217-232.

Rees, J. R., Blazeby, J. M., Fayers, P., Friend, E. A., Welsh, F. K., John, T. G., & Rees, M. (2012). Patient-reported outcomes after hepatic resection of colorectal cancer metastases. *J Clin Oncol, 30*(12), 1364-1370. doi: 10.1200/JCO.2011.38.6177

Rosenberg, W., Donald, A. (1995). Evidence based medicine: An approach to clinical problem-solving. *British Medical Journal, 1122*.

Stewart, A., Greenfield, S., Hays, R., Wells, K., Rogers, W., Berry., S., McGlynn, E., Ware. J. (1989). Functional status and well-being of patients with chronic condistions. Results from the Medical Outcomes Study. *JAMA, 262*(7).

Tan, M. L., Idris, D.B., Teo, L.W., Loh, S.Y., Seow, G.C., Chia, Y.Y., Tin, A.S. (2014). Validation of EORTC-QLQ-C30 and QLQ-BR23 questionnaires in the measurement of quality of life of breast cancer in Singapore. *Asia-Pacific Journal of Oncology Nursing, 1*(1), 22-32.

Terwee, C. B., Dekker, F.W., Wiersinga, W.M., Prummel, M.F., Bossuyt, M.M. (2003). On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. *Quality of Life Research, 12*(4), 349-362.

Trochim, W. (2006). The Multitrait-Multimethod Matrix. from http://www.socialresearchmethods.net/kb/mtmmmat.php

von Karsa, L., Patnick, J., Segnan, N., Atkin, W., Halloran, S., Lansdorp-Vogelaar, I., . . . Valori, R. (2013). European guidelines for quality assurance in colorectal cancer screening and diagnosis: overview and introduction to the full supplement publication. *Endoscopy, 45*(1), 51-59. doi: 10.1055/s-0032-1325997

Wan, C., Meng, Q., Yang, Z., Tu, X., Feng, C., Tang, X., Zhang, C. (2008). Validiation of the simplified Chinese version of the EORTC-QLQ-C30 from the measurements of five types of inpatients with cancer. *Annals of Oncology, 19*, 2053-2060.

Ware, J. E., Gandek, B. (1998). Overview of the SF-36 health survey and the International Quality of Life Assessment (IQOLA) Project. *J Clin Epidemiol, 51*(11), 903-912.

Ware, J. E., Sherbourne, C. (1992). The MOS 36-Item Short Form Health Survey (SF-36). *Medical Care, 30*(6), 473 - 483.

Ware, J. J. (1984). Methodology in behavioral and psychosocial cancer research. Conceptualizing disease impact and treatment outcomes. *Cancer, 53*, 2316-2326.

Ware, J. J., Brook, R.H., Davies-Avery, A. (1980). Conceptualization and measurement of health of adults in the Health Insurance Study. *Model of Health and Methodology, 1*.

WHO. (1948). *Constitution of the World Health Organization*. Paper presented at the International Health Conference, New York.

Wilson, I. B., Cleary, P. (1995). Linking Clinical Variables with Health-Related Quality-of-Life: A Conceptual Model of Patient Outcomes. *JAMA, 273*(1), 59-65.

Appendices

Appendix I - Questionnaires

English version of the EORTC-QLQ-LM21

EORTC QLQ – LMC21

Patients sometimes report that they have the following symptoms or problems. Please indicate the extent to which you have experienced these symptoms or problems during the past week. Please answer by circling the number that best applies to you.

During the past week :	**Not at All**	**A Little**	**Quite a Bit**	**Very Much**
31. Have you had trouble with eating?	1	2	3	4
32. Have you felt full up too quickly after beginning to eat?	1	2	3	4
33. Have you worried about losing weight?	1	2	3	4
34. Have you had problems with your sense of taste?	1	2	3	4
35. Have you had a dry mouth?	1	2	3	4
36. Have you had a sore mouth or tongue?	1	2	3	4
37. Have you been less active than you would like to be?	1	2	3	4
38. Have you had tingling hands or feet?	1	2	3	4
39. Have you had pain in your stomach area?	1	2	3	4
40. Have you had discomfort in your stomach area?	1	2	3	4
41. Have your skin or eyes been yellow (jaundiced)?	1	2	3	4
42. Have you had pain in your back?	1	2	3	4
43. Have you felt slowed down?	1	2	3	4
44. Have you felt lacking in energy?	1	2	3	4
45. Have you had trouble having social contact with friends?	1	2	3	4
46. Have you had trouble talking about your feelings to your family or friends?	1	2	3	4
47. Have you felt stressed?	1	2	3	4
48. Have you felt less able to enjoy yourself?	1	2	3	4
49. Have you worried about your health in the future?	1	2	3	4
50. Were you worried about your family in the future?	1	2	3	4
During the past four weeks:				
51. Has the disease or treatment affected your sex life (for the worse)?	1	2	3	4

Pre-translated EORTC-QLQ-LMC21

NORWEGIAN

EORTC QLQ – LMC21

En del pasienter opplever av og til at de har noen av følgende symptomer eller problemer. Vær vennlig å angi i hvilken grad du har hatt disse symptomene eller problemene i løpet av den siste uka. Sett en ring rundt det tallet som best beskriver din tilstand.

I løpet av den siste uka:	Ikke i det hele tatt	Litt	En del	Svært mye
31. Har det vært vanskelig å spise?	1	2	3	4
32. Har du følt deg mett for fort når du spiser?	1	2	3	4
33. Have you worried about losing weight?	1	2	3	4
34. Har du hatt problemer med smakssansen?	1	2	3	4
35. Har du vært tørr i munnen?	1	2	3	4
36. Har du vært sår i munnen eller på tungen ?	1	2	3	4
37. Har du vært mindre aktiv enn ønskelig?	1	2	3	4
38. Har du hatt kribling i hender eller føtter?	1	2	3	4
39. Har du hatt smerter i mageregionen?	1	2	3	4
40. Har du følt ubehag i mageregionen?	1	2	3	4
41. Have your skin or eyes been yellow (jaundiced)?	1	2	3	4
42. Har du hatt vondt i ryggen?	1	2	3	4
43. Har du følt deg redusert?	1	2	3	4
44. Have you felt lacking in energy?	1	2	3	4
45. Har du hatt vanskeligheter med å ha sosial omgang med venner?	1	2	3	4
46. Have you had trouble talking about your feelings to your family or friends?	1	2	3	4
47. Have you felt stressed?	1	2	3	4
48. Have you felt less able to enjoy yourself?	1	2	3	4
49. Har du vært engstelig for helsen din i fremtiden?	1	2	3	4
50. Were you worried about your family in the future?	1	2	3	4
I løpet av de siste 4 ukene:				
51. Has the disease or treatment affected your sex life (for the worse)?	1	2	3	4

NORWEGIAN

EORTC QLQ – LMC21

En del pasienter opplever av og til at de har noen av følgende symptomer eller problemer. Vær vennlig å angi i hvilken grad du har hatt disse symptomene eller problemene i løpet av den siste uka. Sett en ring rundt det tallet som best beskriver din tilstand.

I løpet av den siste uka:	Ikke i det hele tatt	Litt	En del	Svært mye
31. Har det vært vanskelig å spise?	1	2	3	4
32. Har du følt deg mett for fort når du spiser?	1	2	3	4
33. Har du bekymret deg for vekttap?	1	2	3	4
34. Har du hatt problemer med smakssansen?	1	2	3	4
35. Har du vært tørr i munnen?	1	2	3	4
36. Har du vært sår i munnen eller på tungen?	1	2	3	4
37. Har du vært mindre aktiv enn ønskelig?	1	2	3	4
38. Har du hatt kribling i hender eller føtter?	1	2	3	4
39. Har du hatt smerter i mageregionen?	1	2	3	4
40. Har du følt ubehag i mageregionen?	1	2	3	4
41. Har huden eller øyene dine vært gulaktige (gulsott)?	1	2	3	4
42. Har du hatt vondt i ryggen?	1	2	3	4
43. Har du følt deg redusert?	1	2	3	4
44. Har du følt at du mangler energi?	1	2	3	4
45. Har du hatt vanskeligheter med å ha sosial omgang med venner?	1	2	3	4
46. Har du hatt vanskeligheter med å snakke om følelsene dine med familie eller venner?	1	2	3	4
47. Har du følt deg stresset?	1	2	3	4
48. Har du følt at du er mindre i stand til å more deg?	1	2	3	4
49. Har du vært bekymret for din fremtidige helsetilstand?	1	2	3	4
50. Var du bekymret for din familie i fremtiden?	1	2	3	4
I løpet av de siste 4 ukene:				
51. Har sykdommen eller behandlingen påvirket sexlivet ditt (negativt)?	1	2	3	4

NORWEGIAN

EORTC QLQ-C30 (versjon 3.0.)

Vi er interessert i forhold vedrørende deg og din helse. Vær så vennlig å besvare hvert spørsmål ved å sette en ring rundt det tallet som best beskriver din tilstand. Det er ingen "riktige" eller "gale" svar. Alle opplysningene vil bli behandlet konfidensielt.

Ditt navns forbokstaver: ⌊_⌊_⌊_⌊_⌋

Født (dag, mnd, år): ⌊_⌊_⌋_⌊_⌋_⌊_⌊_⌊_⌋

Dato (dag, mnd, år): 31⌊_⌊_⌋_⌊_⌋_⌊_⌊_⌊_⌋

		Ikke i det hele tatt	Litt	En del	Svært mye
1.	Har du vanskeligheter med å utføre anstrengende aktiviteter, slik som å bære en tung handlekurv eller en koffert?	1	2	3	4
2.	Har du vanskeligheter med å gå en lang tur?	1	2	3	4
3.	Har du vanskeligheter med å gå en kort tur utendørs?	1	2	3	4
4.	Er du nødt til å ligge til sengs eller sitte i en stol i løpet av dagen?	1	2	3	4
5.	Trenger du hjelp til å spise, kle på deg, vaske deg eller gå på toalettet?	1	2	3	4

I løpet av den siste uken:

		Ikke i det hele tatt	Litt	En del	Svært mye
6.	Har du hatt redusert evne til å arbeide eller utføre andre daglige aktiviteter?	1	2	3	4
7.	Har du hatt redusert evne til å utføre dine hobbyer eller andre fritidsaktiviteter?	1	2	3	4
8.	Har du vært tung i pusten?	1	2	3	4
9.	Har du hatt smerter?	1	2	3	4
10.	Har du hatt behov for å hvile?	1	2	3	4
11.	Har du hatt søvnproblemer?	1	2	3	4
12.	Har du følt deg slapp?	1	2	3	4
13.	Har du hatt dårlig matlyst?	1	2	3	4
14.	Har du vært kvalm?	1	2	3	4

Bla om til neste side

I løpet av den siste uken:	**Ikke i det hele tatt**	**Litt**	**En del**	**Svært mye**
15. Har du kastet opp?	1	2	3	4
16. Har du hatt treg mage?	1	2	3	4
17. Har du hatt løs mage?	1	2	3	4
18. Har du følt deg trett?	1	2	3	4
19. Har smerter påvirket dine daglige aktiviteter?	1	2	3	4
20. Har du hatt problemer med å konsentrere deg, f.eks. med å lese en avis eller se på TV?	1	2	3	4
21. Har du følt deg anspent?	1	2	3	4
22. Har du vært engstelig?	1	2	3	4
23. Har du følt deg irritabel?	1	2	3	4
24. Har du følt deg deprimert?	1	2	3	4
25. Har du hatt problemer med å huske ting?	1	2	3	4
26. Har din fysiske tilstand eller medisinske behandling påvirket ditt familieliv?	1	2	3	4
27. Har din fysiske tilstand eller medisinske behandling påvirket dine sosiale aktiviteter?	1	2	3	4
28. Har din fysiske tilstand eller medisinske behandling gitt deg økonomiske problemer?	1	2	3	4

Som svar på de neste spørsmålene, sett en ring rundt det tallet fra 1 til 7 som best beskriver din tilstand

29. Hvordan har din helse vært i løpet av den siste uken?

1	2	3	4	5	6	7
Svært dårlig						Helt utmerket

30. Hvordan har livskvaliteten din vært i løpet av den siste uken?

1	2	3	4	5	6	7
Svært dårlig						Helt utmerket

Din Helse og Trivsel

Dette spørreskjemaet handler om hvordan du ser på din egen helse. Disse opplysningene vil hjelpe oss til å få vite hvordan du har det og hvordan du er i stand til å utføre dine daglige gjøremål. ***Takk for at du fyller ut dette spørreskjemaet!***

For hvert av de følgende spørsmålene vennligst sett et ☒ i den ene luken som best beskriver ditt svar.

1. Stort sett, vil du si at din helse er:

Utmerket	Meget god	God	Nokså god	Dårlig
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

2. <u>Sammenlignet med for ett år siden</u>, hvordan vil du si at din helse stort sett er <u>nå</u>?

Mye bedre nå enn for ett år siden	Litt bedre nå enn for ett år siden	Omtrent den samme som for ett år siden	Litt dårligere nå enn for ett år siden	Mye dårligere nå enn for ett år siden
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

3 De neste spørsmålene handler om aktiviteter som du kanskje utfører i løpet av en vanlig dag. Er din helse slik at den begrenser deg i utførelsen av disse aktivitetene nå? Hvis ja, hvor mye?

	Ja, begrenser meg mye	Ja, begrenser meg litt	Nei, begrenser meg ikke i det hele tatt
a Anstrengende aktiviteter som å løpe, løfte tunge gjenstander, delta i anstrengende idrett	☐ 1	☐ 2	☐ 3
b Moderate aktiviteter som å flytte et bord, støvsuge, gå en tur eller drive med hagearbeid	☐ 1	☐ 2	☐ 3
c Løfte eller bære en handlekurv	☐ 1	☐ 2	☐ 3
d Gå opp trappen flere etasjer	☐ 1	☐ 2	☐ 3
e Gå opp trappen én etasje	☐ 1	☐ 2	☐ 3
f Bøye deg eller sitte på huk	☐ 1	☐ 2	☐ 3
g Gå mer enn to kilometer	☐ 1	☐ 2	☐ 3
h Gå noen hundre meter	☐ 1	☐ 2	☐ 3
i Gå hundre meter	☐ 1	☐ 2	☐ 3
j Vaske eller kle på deg	☐ 1	☐ 2	☐ 3

4. I løpet av de siste 4 ukene, hvor ofte har du hatt noen av de følgende problemer i ditt arbeid eller i andre av dine daglige gjøremål på grunn av din fysiske helse?

	Hele tiden	Mye av tiden	En del av tiden	Litt av tiden	Ikke i det hele tatt
a Du har måttet redusere tiden du har brukt på arbeid eller på andre gjøremål	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
b Du har utrettet mindre enn du hadde ønsket	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
c Du har vært hindret i å utføre visse typer arbeid eller gjøremål	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
d Du har hatt problemer med å gjennomføre arbeidet eller andre gjøremål (f.eks. det krevde ekstra anstrengelser)	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

5. I løpet av de siste 4 ukene, hvor ofte har du hatt noen av de følgende problemer i ditt arbeid eller i andre av dine daglige gjøremål på grunn av følelsesmessige problemer (som f.eks. å være deprimert eller engstelig)?

	Hele tiden	Mye av tiden	En del av tiden	Litt av tiden	Ikke i det hele tatt
a Du har måttet redusere tiden du har brukt på arbeid eller på andre gjøremål	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
b Du har utrettet mindre enn du hadde ønsket	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
c Du har utførte arbeidet eller andre gjøremål mindre grundig enn vanlig	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

6. I løpet av <u>de siste 4 ukene</u>, i hvilken grad har din fysiske helse eller følelsesmessige problemer hatt innvirkning på din vanlige sosiale omgang med familie, venner, naboer eller foreninger?

Ikke i det hele tatt	Litt	En del	Mye	Svært mye
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

7. Hvor sterke kroppslige smerter har du hatt i løpet av <u>de siste 4 ukene</u>?

Ingen	Meget svake	Svake	Moderate	Sterke	Meget sterke
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6

8. I løpet av <u>de siste 4 ukene</u>, hvor mye har smerter påvirket ditt vanlige arbeid (gjelder både arbeid utenfor hjemmet og husarbeid)?

Ikke i det hele tatt	Litt	En del	Mye	Svært mye
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

(IQOLA SF-36v2 Standard, Norway (Norwegian))

9. Disse spørsmålene handler om hvordan du har følt deg og hvordan du har hatt det de siste 4 ukene. For hvert spørsmål, vennligst velg det svaralternativet som best beskriver hvordan du har hatt det. Hvor ofte i løpet av de siste 4 ukene har du…

	Hele tiden	Mye av tiden	En del av tiden	Litt av tiden	Ikke i det hele tatt
a Følt deg full av liv?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
b Følt deg veldig nervøs?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
c Vært så langt nede at ingenting har kunnet muntre deg opp?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
d Følt deg rolig og harmonisk?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
e Hatt mye overskudd?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
f Følt deg nedfor og deprimert?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
g Følt deg sliten?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
h Følt deg glad?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
i Følt deg trett?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

10. I løpet av de siste 4 ukene, hvor ofte har din fysiske helse eller følelsesmessige problemer påvirket din sosiale omgang (som det å besøke venner, slektninger osv.)?

Hele tiden	Mye av tiden	En del av tiden	Litt av tiden	Ikke i det hele tatt
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

(IQOLA SF-36v2 Standard, Norway (Norwegian))

11. Hvor RIKTIG eller GAL er <u>hver</u> av de følgende påstander for deg?

	Helt riktig	Delvis riktig	Vet ikke	Delvis gal	Helt gal
a Det virker som om jeg blir syk litt lettere enn andre	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
b Jeg er like frisk som de fleste jeg kjenner	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
c Jeg tror at helsen min vil forverres	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
d Jeg har utmerket helse	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

Takk for at du fylte ut dette spørreskjemaet!

Appendix II – Translation supporting material

Table 14. QLQ-LMC21 Forward translation process

Original English Text	FW1	FW2	FW12
33. Have you worried about losing weight?	Har du engstet deg for vekttap?	Har du bekymret deg for å gå ned i vekt?	Har du bekymret deg for å gå ned i vekt?
41. Have your skin or eyes been yellow (jaundiced)?	Har huden eller øynene dine vært gulaktige (gulsot)?	Har huden eller øynene dine vært gult (gulsot)?	Har huden eller øynene dine vært gulaktige (gulsot)?
44. Have you felt lacking in energy?	Har du følt at du har lite energi?	Har du følt at du mangler energi?	Har du følt at du mangler energi?
46. Have you had trouble talking about your feelings to your family or friends?	Synes du det har vært vanskelig å snakke om dine følelser med familie og venner?	Har du hatt vanskeligheter med å snakke om dine følelser med familie eller venner?	Har du hatt vanskeligheter med å snakke om dine følelser med familie eller venner?
47. Have you felt stressed?	Har du følt deg stresset?	Har du følt deg stresset?	Har du følt deg stresset?
48. Have you felt less able to enjoy yourself?	Har du følt at du er mindre i stand til å more deg?	Har du følt deg mindre i stand til å nyte deg selv?	Har du følt at du er mindre i stand til å more deg?
50. Were you worried about your family in the future?	Bekymrer du deg for din familie i fremtiden?	Har du vært bekymret om din families fremtid?	Har du vært bekymret for din families fremtid?
51. Has the disease or treatment affected your sex life (for the worse)?	Har sykdommen eller behandlingen påvirket sexlivet ditt på en negativ måte?	Har sykdommen eller behandlingen påvirket din sexliv (til det verre)?	Har sykdommen eller behandlingen påvirket sexlivet ditt på en negativ måte?

Table 15. QLQ-LMC21 Backward translation process

First Intermediary Version	Original English Text	BW1	BW2
Har du bekymret deg for å gå ned i vekt?	33. Have you worried about losing weight?	Have you been worried about losing weight?	Have you been worried about losing weight?
Har huden eller øynene dine vært gulaktige (gulsot)?	41. Have your skin or eyes been yellow (jaundiced)?	Have your eyes or skin been yellow-tinted (jaundice)?	Have your skin or your eyes been yellow (jaundiced)?
Har du følt at du mangler energi?	44. Have you felt lacking in energy?	Have you felt at a loss of energy?	Have you experienced a lack of energy?
Har du hatt vanskeligheter med å snakke om dine følelser med familie eller venner?	46. Have you had trouble talking about your feelings to your family or friends?	Have you had difficulties speaking about your feelings with your family or friends?	Have you had difficulties talking about your feelings with family or friends?
Har du følt deg stresset?	47. Have you felt stressed?	Have you felt stressed?	Have you felt stressed?
Har du følt at du er mindre i stand til å more deg?	48. Have you felt less able to enjoy yourself?	Have you felt as if you are less able to have fun?	Have you felt that you are less capable of enjoying yourself?
Har du vært bekymret for din families fremtid?	50. Were you worried about your family in the future?	Have you been worried about your family's future?	Have you been worried about your family's future?
Har sykdommen eller behandlingen påvirket sexlivet ditt på en negativ måte?	51. Has the disease or treatment affected your sex life (for the worse)?	Has your illness of treatment affected your sexlife negatively?	Has the disease or the treatment negatively affected your sex life?

Appendix III - Assessment supporting materials

Table 16. Comparable scales of the SF-36, QLQ-C30 and QLQ-LMC21

SF-36 Bodily Pain	
7	Hvor sterke kroppslige smerter har du hatt I løpet av de siste 4 ukene?
8	I løpet av de siste 4 ukene, hvor mye har smerter påvirket ditt vanlige arbeid (gjelder både arbeid utenfor hjemmet og husarbeid?
QLQ-C30 Pain	
I løpet av den sisten uken:	
9	Har du hatt smerter?
19	Har smerter påvirket dine daglige aktiviteter?
QLQ-LMC21 Abdominal Pain	
I løpet av den siste uka:	
39	Har du hatt smerter I mageregionen?
40	Har du følt ubehag I mageregionen?
42	Har du følt vondt I ryggen?

SF-36 Vitality	
Hvor ofte I løpet av de siste 4 ukene:	
9a	Har du følt deg full av liv?
9e	Har du hatt mye overskudd?
9g	Har du følt deg sliten?
9i	Har du følt deg trett?
QLQ-C30 Fatigue	
I løpet av den sisten uken:	
10	Har du hatt behov for å hvile?
12	Har du følt deg slapp?
18	Har du følt deg trett?
QLQ-LMC21 Activity/vigor	
I løpet av den siste uka:	
37	Har du vært mindre aktiv enn ønskelig?
43	Har du følt deg redusert?
44	Har du følt at du mangler energi?

SF-36 Mental Health	
Hvor ofte I løpet av de siste 4 ukene:	
9b	Følt deg nervøs?
9c	Vært så langt nede at ingenting kar kunnet muntre deg opp?
9d	Følt deg rolig og harmonisk?
9f	Følt deg nedenfor og deprimert?
9h	Følt deg glad?
QLQ-C30 Role - Emotional	
I løpet av den sisten uken:	
21	Har du følt deg anspent?
22	Har du vært engstelig?
23	Har du følt deg irritabel?
24	Har du følt deg deprimert?
QLQ-LMC 21Anxiety	
I løpet av den siste uka:	
47	Har du følt deg stresset?
48	Har du følt at du er mindre I stand til å more deg?
49	Har du vært bekymret for din fremtidige helsestand?
50	Var du bekymret for din familie I fremtiden?

Table 17. Distribution of responses in each category for all items of the QLQ-LMC21 (n=22)

Scale	Item	Response categories				Missing
		1 -ikke i hele i tatt	2 - Litt	3 - En del	4 - Svært mye	
Symptom scales						
Abdominal Pain	39	59.1 %	27.3 %	9.1 %	4.5 %	0.0 %
	40	54.5 %	31.8 %	9.1 %	4.5 %	0.0 %
	42	77.3 %	13.6 %	0.0 %	9.1 %	0.0 %
Activity/Vigor	37	13.6 %	50.0 %	13.6 %	18.2 %	4.5 %
	43	18.2 %	54.5 %	13.6 %	13.6 %	0.0 %
	44[a]	18.2 %	45.5 %	27.3 %	9.1 %	0.0 %
Eating Problems	31	81.8 %	9.1 %	4.5 %	4.5 %	0.0 %
	32	68.2 %	18.2 %	9.1 %	4.5 %	0.0 %
Anxiety	47[a]	81.8 %	18.2 %	0.0 %	0.0 %	0.0 %
	48[a]	59.1 %	31.8 %	4.5 %	4.5 %	0.0 %
	49	18.2 %	59.1 %	13.6 %	9.1 %	0.0 %
	50[a]	45.5 %	36.4 %	9.1 %	9.1 %	0.0 %
Symptom Single Items						
Nutritional Issues	33[a]	72.7 %	22.7 %	4.5 %	0.0 %	0.0 %
Taste Problems	34	68.2 %	18.2 %	13.6 %	0.0 %	0.0 %
Dry Mouth	35	36.4 %	40.9 %	13.6 %	4.5 %	4.5 %
Sore Mouth	36	86.4 %	0.0 %	9.1 %	4.5 %	0.0 %
Peripheral Neuropathy	38	50.0 %	40.9 %	9.1 %	0.0 %	0.0 %
Jaundice	41[a]	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Contact with friends	45	72.7 %	13.6 %	13.6 %	0.0 %	0.0 %
Talking about feelings	46[a]	86.4 %	13.6 %	0.0 %	0.0 %	0.0 %
Sexual Function	51[a]	31.8 %	13.6 %	27.3 %	13.6 %	13.6 %

[a] translated item

Table 18. QLQ-LMC21 Item/scale correlations corrected for overlap

Correlations

	LMC21 39	LMC21 40	LMC21 42	LMC21 37	LMC21 43	LMC21 44	LMC21 31	LMC21 32	LMC21 47	LMC21 48	LMC21 49	LMC21 50	Abdominal Pain 39 (removed 39)	Abdominal Pain 40 (removed 40)	Abdominal Pain 42 (removed 42)	Activity/Vigor 37 (removed 37)	Activity/Vigor 43 (removed 43)	Activity/Vigor 44 (removed 44)	Eating problems 31 (removed 31)	Eating problems 32 (removed 32)	Anxiety 47 (removed 47)	Anxiety 48 (removed 48)	Anxiety 49 (removed 49)	Anxiety 50 (removed 50)
LMC21 39	1	.969**	-.081	.320	.426*	.281	.419	.227	-.193	.133	.015	-.095	.579**	.653**	.992**	.387	.311	.363	.227	.419	.013	-.087	-.037	.027
LMC21 40	.969**	1	.079	.419	.537**	.330	.399	.196	-.078	.306	.141	.032	.712**	.752**	.992**	.475*	.400	.477*	.196	.399	.181	.066	.148	.210
LMC21 42	-.081	.079	1	.332	.452*	.508*	-.125	.153	.181	.465*	.112	.144	.756**	.702**	-.001	.523*	.430*	.371	.153	-.125	.277	.170	.352	.339
LMC21 37	.320	.419	.332	1	.746**	.468*	.309	.482*	-.195	.430	.533*	.541*	.512*	.486*	.372	.667**	.870**	.937**	.482*	.309	.595**	.497*	.469*	.448*
LMC21 43	.426*	.537**	.452*	.746**	1	.681**	.490*	.451*	.012	.598**	.453*	.318	.671**	.648**	.485*	.920**	.753**	.862**	.451*	.490*	.534*	.386	.477*	.544**
LMC21 44	.281	.330	.508*	.468*	.681**	1	.421	.440*	-.149	.319	-.053	-.051	.574**	.587**	.308	.913**	.768**	.550**	.440*	.421	.074	-.085	.089	.096
LMC21 31	.419	.399	-.125	.309	.490*	.421	1	.746**	-.197	-.062	.150	-.174	.174	.204	.413	.498*	.351	.363	.746**	1.000**	-.042	-.067	-.178	-.002
LMC21 32	.227	.196	.153	.482*	.451*	.440*	.746**	1	-.281	-.277	.100	-.116	.236	.278	.213	.486*	.486*	.455*	1.000**	.746**	-.115	-.078	-.268	-.158
LMC21 47	-.193	-.078	.181	-.195	.012	-.149	-.197	-.281	1	.274	.355	.092	.076	.000	-.136	-.073	-.165	-.068	-.281	-.197	.277	.437*	.424*	.575**
LMC21 48	.133	.306	.465*	.430	.598**	.319	-.062	-.277	.274	1	.454*	.508*	.528*	.448*	.221	.503*	.390	.505*	-.277	-.062	.765**	.547**	.844**	.813**
LMC21 49	.015	.141	.112	.533*	.453*	-.053	.150	.100	.355	.454*	1	.688**	.171	.096	.078	.224	.299	.534*	.100	.150	.853**	.917**	.703**	.845**
LMC21 50	-.095	.032	.144	.541*	.318	-.051	-.174	-.116	.092	.508*	.688**	1	.123	.042	-.032	.150	.348	.507*	-.116	-.174	.891**	.881**	.847**	.642**

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 19. EORTC-QLQ-LMC21 Item/scale correlations

Correlations

	LMC21 39	LMC21 40	LMC21 42	LMC21 37	LMC21 43	LMC21 44	LMC21 31	LMC21 32	LMC21 47	LMC21 48	LMC21 49	LMC21 50	LMC21 Abdominal Pain	LMC21 Activity/vigor	LMC21 Eating Problems	LMC21 Anxiety
LMC21 39	1	.969**	-.081	.320	.426*	.281	.419	.227	-.193	.133	.015	-.095	.836**	.403	.341	-.021
LMC21 40	.969**	1	.079	.419	.537**	.330	.399	.196	-.078	.306	.141	.032	.912**	.491*	.313	.156
LMC21 42	-.081	.079	1	.332	.452*	.508*	-.125	.153	.181	.465*	.112	.144	.473*	.476*	.022	.291
LMC21 37	.320	.419	.332	1	.746**	.468*	.309	.482*	-.195	.430	.533*	.541*	.488*	.854**	.429	.526*
LMC21 43	.426*	.537**	.452*	.746**	1	.681**	.490*	.451*	.012	.598**	.453*	.318	.642**	.932**	.503*	.503*
LMC21 44	.281	.330	.508*	.468*	.681**	1	.421	.440*	-.149	.319	-.053	-.051	.512*	.819**	.461*	.044
LMC21 31	.419	.399	-.125	.309	.490*	.421	1	.746**	-.197	-.062	.150	-.174	.304	.484*	.928**	-.073
LMC21 32	.227	.196	.153	.482*	.451*	.440*	.746**	1	-.281	-.277	.100	-.116	.260	.545**	.941**	-.155
LMC21 47	-.193	-.078	.181	-.195	.012	-.149	-.197	-.281	1	.274	.355	.092	-.034	-.149	-.258	.431*
LMC21 48	.133	.306	.465*	.430	.598**	.319	-.062	-.277	.274	1	.454*	.508*	.415	.501*	-.187	.765**
LMC21 49	.015	.141	.112	.533*	.453*	-.053	.150	.100	.355	.454*	1	.688**	.122	.344	.132	.862**
LMC21 50	-.095	.032	.144	.541*	.318	-.051	-.174	-.116	.092	.508*	.688**	1	.040	.280	-.153	.852**
LMC21 Abdominal Pain	.836**	.912**	.473*	.488*	.642**	.512*	.304	.260	-.034	.415	.122	.040	1	.623**	.301	.198
LMC21 Activity/vigor	.403	.491*	.476*	.854**	.932**	.819**	.484*	.545**	-.149	.501*	.344	.280	.623**	1	.552**	.387
LMC21 Eating Problems	.341	.313	.022	.429	.503*	.461*	.928**	.941**	-.258	-.187	.132	-.153	.301	.552**	1	-.124
LMC21 Anxiety	-.021	.156	.291	.526*	.503*	.044	-.073	-.155	.431*	.765**	.862**	.852**	.198	.387	-.124	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 20. Correlations between complementary scales of QLQ-LMC21, QLQ-C30 and SF-36

Correlations

	SF-36 Bodily Pain	SF-36 Vitality	SF-36 Mental Health	C30 Pain	C30 Fatigue	C30 Role Functioning - Emotional	LMC21 Abdominal Pain	LMC21 Activity/vigor	LMC21 Anxiety
SF-36 Bodily Pain	1	.569**	.677**	-.862**	-,327	.468*	-.724**	-.519*	-.548**
SF-36 Vitality	.569**	1	,284	-.467*	-.775**	,097	-.534*	-.798**	-,249
SF-36 Mental Health	.677**	,284	1	-.682**	-,165	.910**	-,272	-,328	-.768**
C30 Pain	-.862**	-.467*	-.682**	1	,309	-.555**	.658**	.507*	.556**
C30 Fatigue	-,327	-.775**	-,165	,309	1	-,032	.522*	.849**	,177
C30 Role	.468*	,097	.910**	-.555**	-,032	1	-,140	-,221	-.786**
LMC21 Abdominal	-.724**	-.534*	-,272	.658**	.522*	-,140	1	.623**	,198
LMC21 Activity/vigor	-.519*	-.798**	-,328	.507*	.849**	-,221	.623**	1	,387
LMC21 Anxiety	-.548**	-,249	-.768**	.556**	,177	-.786**	,198	,387	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Appendix IV – Reliabilities

Table 21. QLQ-LMC21 comparable scale reliability statistics

QLQ-LMC21 Abdominal pain scale reliability statistics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,571	,588	3

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
4,636	3,671	1,9160	3

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	1,545	1,409	1,636	,227	1,161	,014	3

QLQ-LMC21 Vatigue/vigor scale reliability statistics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,835	,835	3

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
6,810	5,862	2,4211	3

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	2,270	2,190	2,381	,190	1,087	,010	3

QLQ-LMC21 Anxiety scale reliability statistics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,735	,723	4

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
6,682	5,370	2,3174	4

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	1,670	1,182	2,136	,955	1,808	,164	4

Table 22. QLQ-C30 comparable scale reliability statistics

QLQ-C30 Pain scale reliability statistics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,833	,836	2

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
3,045	2,807	1,6755	2

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	1,523	1,409	1,636	,227	1,161	,026	2

QLQ-C30 Fatigue scale reliability statistics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,836	,838	3

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
6,048	3,248	1,8021	3

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	2,016	1,952	2,095	,143	1,073	,005	3

QLQ-C30 Emotional role functioning scale reliability statistics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,925	,928	4

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
5,227	5,994	2,4482	4

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	1,307	1,273	1,364	,091	1,071	,002	4

Table 23. SF-36 comparable scale reliability statistics

SF-36 bodily pain scale reliability statistics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,846	,873	2

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
4,409	6,444	2,5384	2

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	2,205	1,864	2,545	,682	1,366	,232	2

SF-36 Vitality scale reliability statistics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,810	,828	4

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
12,955	12,141	3,4843	4

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	3,239	2,818	3,545	,727	1,258	,120	4

SF-36 Mental health scale reliability statistics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,847	,864	5

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
21,818	10,537	3,2460	5

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	4,364	3,773	4,818	1,045	1,277	,252	5

Table 24. QLQ-C30 and QLQ-LMC21 comparable scale reliability statistics

QLQ-C30 and QLQ-LMC21 pain scales combined statistics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,793	,796	5

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
7,682	10,703	3,2716	5

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	1,536	1,409	1,636	,227	1,161	,014	5

QLQ-C30 and QLQ-LMC21 vitality/fatigue combined scale statistics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,884	,891	5

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
10,700	11,274	3,3576	5

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	2,140	2,000	2,350	,350	1,175	,017	5

QLQ-C30 and QLQ-LMC21 emotional role functioning/anxiety combined scale statistics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,904	,912	8

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
11,909	20,277	4,5030	8

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	1,489	1,182	2,136	,955	1,808	,109	8

XX